Université de Sherbrooke

Estimation d'une densité prédictive avec information additionnelle

par

Abdolnasser Sadeghkhani

Thèse présentée au Département de mathématiques en vue de l'obtention
du grade de docteur ès sciences

Faculté des Sciences

Sherbrooke, Québec, Canada
Septembre 2017

Le 12 Septembre 2017

Jury a accepté la thèsse de Monsieur Abdolnasser Sadeghkhani dans sa version finale

Membres du jury

Professeur Éric Marchand
Directeur de recherche
Département de mathématiques
Université de Sherbrooke

Professeur Patrick Richard
Évaluateur interne
Département d'économique
Université de Sherbrooke

Professeur S. Ejaz Ahmed
Évaluateur externe
Department of Mathematics & Statistics
Brock University

Professeur Taoufik Bouezmarni
Président–rapporteur
Département de mathématiques
Université de Sherbrooke

Sommaire

Dans le contexte de la théorie bayésienne et de théorie de la décision, l'estimation d'une densité prédictive d'une variable aléatoire occupe une place importante. Typiquement, dans un cadre paramétrique, il y a présence d'information additionnelle pouvant être interprétée sous forme d'une contrainte. Cette thèse porte sur des stratégies et des améliorations, tenant compte de l'information additionnelle, pour obtenir des densités prédictives efficaces et parfois plus performantes que d'autres données dans la littérature. Les résultats s'appliquent pour des modèles avec données gaussiennes avec ou sans une variance connue. Nous décrivons des densités prédictives bayésiennes pour les coûts Kullback-Leibler, Hellinger, Kullback-Leibler inversé, ainsi que pour des coûts du type $\alpha-$divergence et établissons des liens avec les familles de lois de probabilité du type *skew–normal*. Nous obtenons des résultats de dominance faisant intervenir plusieurs techniques, dont l'expansion de la variance, les fonctions de coût duaux en estimation ponctuelle, l'estimation sous contraintes et l'estimation de Stein. Enfin, nous obtenons un résultat général pour l'estimation bayésienne d'un rapport de deux densités provenant de familles exponentielles.

Université de Sherbrooke

Abstract

Faculté des Sciences

Doctor of Philosophy

by Abdolnasser Sadeghkhani

In the context of Bayesian theory and decision theory, the estimation of a predictive density of a random variable represents an important and challenging problem. Typically, in a parametric framework, usually there exists some additional information that can be interpreted as constraints. This thesis deals with strategies and improvements that take into account the additional information, in order to obtain effective and sometimes better performing predictive densities than others in the literature. The results apply to normal models with a known or unknown variance. We describe Bayesian predictive densities for Kullback–Leibler, Hellinger, reverse Kullback-Leibler losses as well as for α–divergence losses and establish links with skew–normal densities. We obtain dominance results using several techniques, including expansion of variance, dual loss functions in point estimation, restricted parameter space estimation, and Stein estimation. Finally, we obtain a general result for the Bayesian estimator of a ratio of two exponential family densities.

Acknowledgements

There are several individuals who have contributed significantly to my Ph.D. years and the completion of this dissertation. First, I am indebted to my supervisor Prof. Éric Marchand, from whom I have learned as much about research, as about writing, giving talks, as about the importance of providing an amiable ambiance. I would also like to acknowledge my thesis committee members, Professor S. Ejaz Ahmed and Professor Patrick Richard, for their careful reading and helpful comments in spite of lack of time. I would also like to express my appreciation to Professor Taoufik Bouezmarni of Département mathématiques for his support throughout the course of this work. I have been exceedingly fortunate to have many delightful friends among the professors, staff and graduate students. They are the sole reason why the past five years have been so thoroughly enjoyable.

Finally, I am thankful wholeheartedly to my wife, Fatima, for her understanding and for accompanying me.

Contents

Abstract **iii**

Acknowledgements **iv**

List of Figures **viii**

List of Tables **ix**

Abbreviations **x**

Introduction **1**

1 Definitions and preliminaries **6**

1.1 Introduction 6

1.2 Predictive density estimation 7

1.2.1 Improvement by variance expansion 12

1.2.2 Duality between point estimation and predictive density estimation 15

1.3 Point estimation in restricted parameter spaces . . . 16

1.4 Bayes estimators and predictive density estimators in exponential families 20

1.5 Skew–normal and other skewed distributions 22

1.5.1 Generalized Balakrishnan type skew-normal distribution 23

1.5.2 Skew-Student t distribution 26

2 Predictive Density Estimation When the Variance is Known **30**
2.1 Introduction 32
2.1.1 Problem and Model 32
2.1.2 Predictive density estimation 33
2.1.3 Description of main findings 34
2.2 Bayesian predictive density estimators and skewed normal type distributions 36
2.2.1 Bayesian predictive density estimators 36
2.2.2 Examples of Bayesian predictive density estimators 39
2.2.2.1 Univariate case with $\theta_1 \geq \theta_2$ . . 40
2.2.2.2 Univariate case with $|\theta_1 - \theta_2| \leq m$ 42
2.2.2.3 Multivariate case with $||\theta_1 - \theta_2|| \leq m$ 43
2.2.2.4 reverse kullback-leibler loss 44
2.3 General dominance results 45
2.3.1 Improvements by variance expansion 46
2.3.2 Improvements through duality 50
2.4 Bayesian dominance results 55
2.4.1 Reverse Kullback-Leibler loss 56
2.4.2 Kullback-Leibler loss 58
2.5 Examples, illustrations and further comments 62
2.6 Concluding remarks 64

3 Predictive Density Estimation With Unknown Variance **74**
3.1 Introduction 74
3.2 Bayes posterior analysis 75
3.2.1 Posterior distributions and expectations . . . 75
3.2.2 Predictive densities 79
3.3 Improving on plug–in predictive density estimators under KL loss function 84
3.4 Improving on plug–in predictive density estimators under RKL loss function 87
3.5 Concluding remarks 89

4 Density Ratio Estimation **92**
4.1 Introduction 92
4.2 Bayesian density ratio estimation 92
4.3 Examples 94
4.4 Conclusion 96

Conclusion and future work **98**

Bibliography **101**

List of Figures

1.1 Density of $\mathbb{SN}_1(n, 1, 2, 0, 1)$ for $n = 1, 3$ and 10. 25
1.2 Density of $\mathbb{SN}_1(3, \alpha_0, 2, 0, 1)$ for $\alpha_0 = -2$, 0 and 2. . . 25
1.3 Density of $\mathbb{SN}_1(3, 1, \alpha_1, 0, 1)$ for $\alpha_1 = -2, 0$ and 2. . . 25

2.1 Kullback-Leibler risk ratios for $p = 1$, $A = [0, \infty)$, and $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2 = 1$ 63
2.2 Kullback-Leibler risk ratios for $p = 1$, $A = [0, \infty)$, $\sigma_1^2 = \sigma_Y^2 = 1$ and $\sigma_2^2 = 1, 2, 4$ 63
2.3 Kullback-Leibler risk ratios for $p = 1$, $A = [-1, 1]$, and $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2 = 1$ 64
2.4 Kullback-Leibler risk ratios for $p = 1$, $A = [-2, 2]$, and $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2 = 1$ 64

3.1 Risk ratio of the Bayes and MRE predictive density estimator with $k = 3$ and $A = [0, +\infty)$. 90
3.2 Risk ratio of the Bayes and the MRE predictive density estimator with $k = 3$ and $A = [-6, 6]$. 90

List of Tables

1.1 Some densities from model (1.20) with their natural parameters and dual losses. 22

Abbreviations

cdf	**C**umulative **D**ensity **F**unction
H	**H**ellinger loss
DR(E)	**D**ensity **R**atio (**E**stimation)
KL	**K**ullback **L**eibler
ML	**M**aximum **L**ikelihood
MRE	**M**inimum **R**isk **E**quivariant
pdf	**P**robability **D**ensity **F**unction
RNL	**R**eflected **N**ormal **L**oss
RKL	**R**everse **K**ullback **L**eibler
SEL	**S**quared **E**rror **L**oss

In memory of my dear father,
To my wife & my mother…

Introduction

Density estimation, where data from X is used to estimate the density of Y is one of the fundamental problems in statistics. Predictive density estimation involves drawing data from $X \sim p_\theta$ to obtain an estimate of the density q_θ of Y for prediction purposes. Such a density is of interest as a surrogate and for generating either future or missing values of Y.

To assess efficiency and to aid in selecting a predictive density estimator $\hat{q}(\cdot; X)$, we adopt Bayesian and decision theoretic perspectives, thus introducing a loss function based on a divergence measure of the distance between two densities. We will be particularly interested in evaluating predictive density estimators in terms of frequentist risk, as well as normal models with additional information available on θ, as presented below in the Brief outline.

The developments in this thesis relate to statistical inference for restricted parameters and here are some situations where such additional information arises.

Example I. (An application in psychology)
Researchers in the social and behavioral sciences often have clear expectations and beliefs about the order and direction of the parameters in their statistical model. For example, a researcher might expect that one regression coefficient β_1 is larger than others β_2 and

β_3. In such a case, one is interested in either adapting priors to the restriction, or elaborating procedures that capitalize on the information potentially leading to lower frequentist risk. Vanbrabant et al. [1] showed how such a constraint leads to gains in sample size reduction in a hypothesis testing context, and they provided an illustration on the impact of cognitive behavioral therapy to treat depression.

Example II. (Relation between El Niño and hurricanes)
El Niño refers to unusually warm ocean currents in the Pacific that appear around Christmas time. Monsoon rains in the central pacific and droughts and forest fires in Indonesia and Australia have been linked to El Niño. The following hypothesis appears in Kitchens [2]:

H : Warm phases of El Niño suppresses hurricanes, while cold weather fosters tl

Information from 1950 to 1995 was used based on a simple one–way classification, $X_{ij} = \mu_i + e_{ij}$ by admitting that El Niño has 3 levels: cold ($i = 1$), neutral ($i = 2$) and warm ($i = 3$), with X_{ij} representing the number of hurricanes and μ_i its expectation. Note that hypothesis H is equivalent to hypothesis $H' : \mu_1 \geq \mu_2 \geq \mu_3$. If H' assumed to be plausible, then the prediction of future values should exploit the precise information.

Example III. (An application in Biology)
The following example is presented in Liseo and Loperfido [3]. Suppose we observe a random sample replications of a bivariate normal density (X_1, X_2) with mean vector (μ_1, μ_2), where X_1 is the length of the right leg in an adult man while X_2 is the corresponding length of the left leg. Then it is reasonable to assume that $|\mu_1 - \mu_2| \leq c$ for

some constant c. Estimates of μ_1 and μ_2, based on additional information seem more appropriate. The same can be said for obtaining on density estimate of $Y_1 \sim \mathbb{N}(\mu_1, \sigma_Y^2)$. This type of problem is at the heart of the thesis.

B) Brief outline of thesis

Chapter 1 deals with definitions and preliminaries which are needed throughout the thesis, as well as some underlying lemmas and theorems. We study the duality between point estimators and plug–in predictive density estimators related to different loss functions, along with determination of Bayesian predictive densities. We briefly discuss estimation for restricted parameters. Afterwards, some skew–normal, skew–Student t, and skewed type distributions are presented.

In Chapter 2, based on the model

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathbb{N}_{2p}\left(\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \, \Sigma = \left(\begin{smallmatrix} \sigma_1^2 I_p & 0 \\ 0 & \sigma_2^2 I_p \end{smallmatrix}\right)\right), Y_1 \sim \mathbb{N}_p(\theta_1, \, \sigma_Y^2 I_p) \,,$$

with known σ_1^2, σ_2^2, and σ_Y^2, we address the question of providing competitive predictive density estimates for the density of Y_1, in comparison to other available predictive densities, such as plug-in densities, and those obtained by the criteria of maximum likelihood or minimum risk equivariance. Results relate to the class of α-divergence losses. This developments exploit the presence of additional information of the form $\theta_1 - \theta_2 \in A \subset \mathbb{R}^p$, with known A. We investigate how to gain from the additional information in providing a predictive density $\hat{q}(\cdot\,; X)$ as an estimate of the density $q_{\theta_1}(\cdot)$ of Y_1. Indeed, additional information $\theta_1 - \theta_2 \in A$ renders X_2 useful in estimating the density of Y_1 despite the independence and the otherwise unrelated parameters. An ensemble of techniques are exploited, including variance expansion (for KL loss), point estimation duality, and concave inequalities. Representations for Bayesian predictive densities, and in particular for $\hat{q}_{\pi_{U,A}}$ associated with a uniform prior for θ truncated to $\{\theta : \theta_1 - \theta_2 \in A\}$, are established and are used for the Bayesian dominance findings. Finally and interestingly, these Bayesian predictive densities also relate to skew-normal distributions, as well as new forms of such distributions.

Chapter 3 considers the topics in Chapter 2 in the normal set–up with unknown variance. More specifically, for $X_i \sim \mathbb{N}_p(\theta_i, \sigma^2 I_p)$, $i = 1, 2$, independent of $S^2 \sim \sigma^2 \chi^2_k$, we study predictive density estimation of the density of $Y_1 \sim \mathbb{N}_p(\theta_1, \sigma^2 I_p)$ for Kullback–Leibler and reverse Kullback–Leibler losses and $\theta_1 - \theta_2 \in A$. Interesting posterior and predictive density representation arise and we provide improvements on plug–in densities.

Chapter 4 is a note on density ratio estimation. This topic is connected to several problems such as machine learning and has attracted much attention in the literature. We present a general representation for Bayesian ratio estimators under squared–log error loss in the context of exponential family densities. The class of Bayesian estimators are seen to include in some cases ratios of plug–in density estimators. The result is general with respect to the family of densities and choice of priors.

Chapter 1

Definitions and preliminaries

1.1 Introduction

In this thesis, we will focus on improving predictive density estimators under additional parametric information and we present here preliminary related results. Predictive density estimation is briefly reviewed in Section 1.2 with a focus on the notable α–divergence class of loss functions including Kullback–Leibler (KL), reverse Kullback–Leibler (RKL), and Hellinger (H) losses as specific cases. We will be studying the frequentist risk of plug–in type predictive density estimators, and we will expand on duality connections with point estimation, variance expansion improvements, and Bayesian procedures. Section 1.3 deals with point estimation of a multivariate normal mean with additional information, while Section 1.4 looks at point estimation and predictive density estimation for exponential families. Finally, skew–normal, as well as other skewed distributions will be presented in Section 1.5. A rich family of Bayesian predictive density estimators arise in the presence of constraints on the parameters, and these include some of the skew–normal and skew–Student t distributions of Section 1.5.

1.2 Predictive density estimation

Predictive analysis is about extracting information from historical and current data to predict future trends. The statistical prediction of future values of a random variable, based on an observed learning sample, appears in a variety of problems. It has been argued in the literature that prediction, as opposed to parameter estimation, is the proper activity of statisticians, because prediction is often the scientific question of interest and partly because the ability of statisticians to predict can actually be checked. The sampling distribution of Y (possibly a vector with dimension, $p > 1$) given $X = x$ and (a) known parameter(s) would be an obvious predictive distribution, but without knowledge of the underlying parameter, it cannot be used. (see for instance Nayak [4] and the references below.)
Consider the conditionally independent random variables with Lebesgue densities

$$X|\theta \sim p_\theta(x), \quad Y|\theta \sim q_\theta(y), \; x, \, y \in \mathbb{R}^p, \tag{1.1}$$

and $\theta \in \Theta \subseteq \mathbb{R}^p$. The goal is to estimate the future density Y based on X. Given a prior density π for θ with cdf G, the conditional or posterior distribution of Y is given by

$$\begin{aligned} q(y|x) &= \frac{p_\theta(y,x)}{p(x)} \\ &= \int_\Theta \frac{p(y,x,\theta)}{p(x)} \, dG(\theta) \\ &= \int_\Theta p_\theta(y|x) \frac{\pi(\theta)\, p_\theta(x)}{p(x)} \, dG(\theta) \\ &= \int_\Theta p_\theta(y) \pi(\theta|x) \, dG(\theta), \end{aligned} \tag{1.2}$$

given the conditional independence. This induces the Bayesian predictive density estimator as the posterior expectation of $p_\theta(y)$, or a mixture of the $p_\theta(y)$'s. In much of the statistical literature, $q(y|x)$ is referred to the posterior predictive distribution for Y.

Given a predictive density estimate $\hat{q}(\cdot;x)$, $x \in \mathbb{R}^p$, several loss functions are at our disposal for measuring the proximity of $\hat{q}$ to q_θ. These include Kullback-Leibler (KL) loss given by

$$L_{KL}(\theta, \hat{q}(\cdot;\, x)) = \int_{\mathbb{R}^p} q_\theta(y) \, \log \frac{q_\theta(y)}{\hat{q}(y;x)} \, dy, \tag{1.3}$$

as well as reverse Kullback–Leibler (RKL) loss,

$$L_{RKL}(\theta, \hat{q}(\cdot;\, x)) = \int_{\mathbb{R}^p} \hat{q}(y;x) \, \log \frac{\hat{q}(y;x)}{q_\theta(y)} \, dy, \tag{1.4}$$

The loss functions in (1.3) and (1.4) belong to the class of $\alpha-$divergence loss functions (e.g., Csiszàr, [5]) given by

$$L_\alpha(\theta, \hat{q}) \;=\; \int_{\mathbb{R}^p} h_\alpha\left(\frac{\hat{q}(y;x)}{q_\theta(y)}\right) \, q_\theta(y)\, dy\,, \tag{1.5}$$

with

$$h_\alpha(z) = \begin{cases} \frac{4}{1-\alpha^2}(1 - z^{(1+\alpha)/2}) & \text{for } |\alpha| < 1 \\ z\log(z) & \text{for } \alpha = 1 \\ -\log(z) & \text{for } \alpha = -1. \end{cases} \tag{1.6}$$

The KL and RKL losses correspond to $\alpha = -1$ and 1 respectively, while Hellinger loss is associated with $\alpha = 0$. The performance of predictive densities $\hat{q}(\cdot;X)$ related to L_α in (1.5) may be measured by the frequentist risk

$$R_\alpha(\theta, \hat{q}) \;=\; \int_{\mathbb{R}^p} L_\alpha\left(\theta, \hat{q}(\cdot;x)\right) \, p_\theta(x)\, dx\,. \tag{1.7}$$

Predictive density (1.2) is the Bayes estimate for KL loss. The following result provides the Bayes predictive density estimate for the $\alpha-$divergence loss with $-1 \leq \alpha < 1$, while the case of RKL loss is presented in the Appendix of Chapter 2.

Lemma 1.2.1. *In model (1.1), given a prior density π with respect to the σ–finite measure ν with cdf G, the Bayes predictive density estimator $\hat{q}_\pi(\cdot;X)$ of the density of $q_\theta(\cdot)$ under loss function L_α in (1.5) for $\alpha \neq 1$, is given by*

$$\hat{q}_\pi(y;x) \propto \left\{ \int_{\mathbb{R}^p} q_\theta(y)^{\frac{1-\alpha}{2}} \, \pi(\theta|x) \, d\nu(\theta) \right\}^{\frac{2}{1-\alpha}} .$$

Proof. The expected posterior risk of $\hat{q}(\cdot)$ under $\alpha-$divergence loss in (1.5) is equal to

$$\begin{aligned} & \frac{4}{1-\alpha^2}\left(1 - \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \hat{q}(y;x)^{\frac{1+\alpha}{2}} \, q_\theta(y)^{\frac{1-\alpha}{2}} \pi(\theta|x) \, d\nu(\theta) \, dy\right) \\ = & \frac{4}{1-\alpha^2}\left(1 - \int_{\mathbb{R}^p} \hat{q}(y;x)^{\frac{1+\alpha}{2}} k(y;x) \, dy\right), \end{aligned} \tag{1.8}$$

where $k(y;x) = \int_{\mathbb{R}^p} q_\theta(y)^{\frac{1-\alpha}{2}} \pi(\theta|x) \, d\nu(\theta)$. Minimizing (1.8) in $\hat{q}(\cdot)$ is equivalent to maximizing $\int_{\mathbb{R}^p} \hat{q}^{\frac{1+\alpha}{2}} k(y;x) \, dy$. An application of Hölder's inequality $\int f\, g \leq (\int f^a)^{1/a} (\int g^b)^{1/b}$, with $f = k(y;x)$, $g = \hat{q}^{\frac{1+\alpha}{2}}(y;x)$, $a = \frac{2}{1-\alpha}$ and $b = \frac{2}{1+\alpha}$ ($\frac{1}{a} + \frac{1}{b} = 1$), and with equality iff $\hat{q}(y;x) \propto k^{\frac{2}{1-\alpha}}(y;x)$, yields the result. □

Next, we consider conditional independently distributed

$$X \sim \mathrm{N}_p(\theta, \sigma_X^2 \, I_p) \, , \quad Y \sim \mathrm{N}_p(\theta, \sigma_Y^2 \, I_p) \, , \tag{1.9}$$

with common unknown mean θ and known variances σ_X^2 and σ_Y^2. The next example provides the Bayes predictive density estimator associated with α–divergence loss and a normal prior distribution.

Example 1.2.1. *Consider model (1.9) and the normal prior $\pi_0(\theta) \sim \mathrm{N}_p\left(\mu, \tau^2 I_p\right)$ under α–divergence loss in (1.5) with $-1 \leq \alpha < 1$.*

Then, the Bayes predictive density $\hat{q}_{\pi_0}$ is given by

$$\hat{q}_{\pi_0}(\cdot;\, x) \;\sim\; \mathbb{N}_p\left(\gamma\, x + (1-\gamma)\, \mu,\; \left(\frac{(1-\alpha)\, \tau^2 \sigma_X^2}{2\,(\tau^2+\sigma_X^2)} + \sigma_Y^2\right) I_p\right),$$

where $\gamma = \frac{\tau^2}{\sigma_X^2+\tau^2}$.

Proof. According to Lemma 1.2.1, the Bayes predictive density estimate of $q_{\theta_1}(y)$ is given by

$$\begin{aligned}\hat{q}_{\pi_0}(y;x) &\propto \left\{\int_{\mathbb{R}^p} \phi^{(1-\alpha)/2}\left(\frac{\theta-y}{\sigma_Y}\right) \pi_0(\theta|x)\, d\theta\right\}^{2/1-\alpha} \\ &\propto \left\{\int_{\mathbb{R}^p} \phi\left(\frac{y-\theta}{\sqrt{\frac{2}{1-\alpha}}\sigma_Y}\right) \phi\left(\frac{\theta-x}{\sigma_X}\right) \phi\left(\frac{\theta-\mu}{\tau}\right) d\theta\right\}^{2/1-\alpha},\end{aligned}$$

given that $\phi^m(z) \propto \phi(m^{1/2}z)$. By making use of decomposition

$$\frac{\|t-a_1\|^2}{b_1^2} + \frac{\|t-a_2\|^2}{b_2^2} = \frac{\|a_1-a_2\|^2}{b_1^2+b_2^2} + \frac{\|t-c\|^2}{d^2},$$

where, $c = \frac{a_1 b_2^2 + a_2 b_1^2}{b_1^2+b_2^2}$, $d^2 = \frac{b_1^2 b_2^2}{b_1^2+b_2^2}$, we can write

$$\phi\left(\frac{\theta-x}{\sigma_X}\right) \phi\left(\frac{\theta-\mu}{\tau}\right) \;\propto\; \phi\left(\frac{\theta-w}{\sigma_W}\right) \phi\left(\frac{x-\mu}{\sqrt{\sigma_X^2+\tau^2}}\right),$$

$$\phi\left(\frac{\theta-y}{\sqrt{\frac{2}{1-\alpha}}\,\sigma_Y}\right) \phi\left(\frac{\theta-w}{\sigma_W}\right) \;\propto\; \phi\left(\frac{y-u}{\sqrt{\frac{2}{1-\alpha}\,\sigma_Y^2+\sigma_U^2}}\right) \phi\left(\frac{\theta-u}{\sigma_U}\right),$$

with $w = \frac{x\tau^2+\mu\sigma_X^2}{\sigma_X^2+\tau^2}$, $u = \frac{y\sigma_W^2 + \frac{2w}{1-\alpha}\sigma_Y^2}{\sigma_W^2+\frac{w}{1-\alpha}\sigma_Y^2}$, $\sigma_W^2 = \frac{\tau_1^2\,\sigma_1^2}{\tau_1^2+\sigma_1^2}$ and $\sigma_U^2 = \frac{2\,\sigma_Y^2/(1-\alpha)\,\sigma_W^2}{2\,\sigma_Y^2/(1-\alpha)+\sigma_W^2}$. Since $\hat{q}_{\pi_0}(y;x) \propto (\phi(\frac{y-u}{\sqrt{\frac{2}{1-\alpha}\sigma_Y^2+\sigma_U^2}}))^{\frac{2}{1-\alpha}}$, the result follows. □

Remark 1.2.2. *In Example 1.2.1, setting $\tau^2 \to \infty$, yields the Bayes predictive density with respect to the uniform prior $\pi_0(\theta) = 1$, or the minimum risk equivariant (MRE) predictive density estimator. It is* *given by*

$$\hat{q}_{mre}(\cdot, x) \sim \mathbb{N}_p \left(x, \left(\frac{(1-\alpha)\sigma_X^2}{2} + \sigma_Y^2 \right) I_p \right) , \tag{1.10}$$

as obtained by Ghosh et al. [6]. For Kullback–Leibler loss (i.e. $\alpha = -1$*) we obtain*

$$\hat{q}_{mre}(\cdot, x) \sim \mathbb{N}_p \left(x, (\sigma_X^2 + \sigma_Y^2) I_p \right) , \tag{1.11}$$

such as given by Aitchison [7].

Definition 1.2.3 (Dominance). Let $\hat{q}_1$ and $\hat{q}_2$ be two predictive density estimators in estimating q_θ. Then $\hat{q}_1$ dominates $\hat{q}_2$ if we have $R(\theta, \hat{q}_1) \leq R(\theta, \hat{q}_2)$ for all $\theta \in \Theta$, with strictly inequality for some θ.

A plausible approach to find a predictive density estimator is obtained by replacing θ by an estimator $\hat{\theta}(X)$ in the true density $q_\theta(\cdot)$, yielding $q_{\hat{\theta}(X)}(\cdot)$, which is called a *plug–in* predictive density estimator. There are methods available to improve on plug–in predictive density estimators such as variance expansion. See for instance Fourdrinier et al. [8] as well as Section 1.2.1.
The seminal papers of George, Liang, and Xu [9] as well as Brown, George and Xu [10], shed new light on the relationship between predictive density estimation and shrinkage estimation. They showed there exists a duality between predictive density estimation under KL loss and the problem of estimating the mean of the multivariate normal distribution under SEL in model (1.9). They obtained large class of minimax predictive densities, as well as Bayes admissible predictive densities. Their results revolved around the inadmissibility of $\hat{q}_{mre}$ for $p \geq 3$, a result established earlier by Komaki [11].

Also, they showed the generalized Bayes estimator for the mean of a p–variate normal distribution shares many properties like minimaxity, best invariant estimator under location-scale transformation, constant risk, admissibility for $p \le 2$ and inadmissibility for $p \ge 3$. More information on the dual relationship between density estimation under α–divergence and the point estimation problem can be found in Ghosh et al. [6].

1.2.1 Improvement by variance expansion

Consider predictive densities for Y in model (1.1) of the form

$$q_{\hat{\theta},c}(\cdot, x) \sim \mathbb{N}_p(\hat{\theta}(x), c\,\sigma_Y^2 I_p)\,, \tag{1.12}$$

where $\hat{\theta}(x)$ is an estimate of θ. Cases $c = 1$, correspond to plug-in predictive density estimators, while cases $c > 1$ correspond to *scale expanded variants*. Previous work (e.g., Aitchison [7], Fourdrinier et al., [8]; Kubokawa, Marchand and Strawderman, [12], [13]) have shown that such scale expansions are interesting and can provide significant risk improvement on plug-in procedures. The MRE predictive estimator in (1.10) related to L_α is obtained by taking $c = 1 + \frac{(1-\alpha)\,\sigma_X^2}{2\,\sigma_Y^2}$ and $\hat{\theta}(X) = X$. The following theorem elaborates on a general phenomenon concerning plug-in estimators and how they can be improved upon within the class of normal density estimators in (1.12).

Theorem 1.2.4. *Consider model (1.9) and $\theta \in C$. Let $\delta(X)$ be an estimator of θ, with risk $R(\theta, \delta) = E\|\delta(X) - \theta\|^2$, and $\underline{R} = \inf_{\theta \in C} R(\theta, \delta) > 0$. For estimating the density of $Y \sim \mathbb{N}_p(\theta, \sigma_Y^2\, I_p)$, the predictive density estimator $\hat{q}_c \sim \mathbb{N}_p(\delta(X), c\,\sigma_Y^2\, I_p)$ dominates $\hat{q}_1 \sim N_p(\delta(X), \sigma_Y^2\, I_p)$ under KL loss if $1 < c \le (1 + \frac{\underline{R}}{p\,\sigma_Y^2})$, and iff*

$1 < c \leq c_0(1+\frac{R}{p\,\sigma_Y^2})$, *with* $c_0(m)$ *the root of* $G_m(c) = (1-c^{-1})m - \log c$ *on* (m, ∞).

Proof. The difference in risks is equal to

$$\begin{aligned}
R_{KL}(\theta, \hat{q}_1) - R_{KL}(\theta, \hat{q}_c) &= \mathbb{E}^{X,Y}\left[\log\left(\frac{\hat{q}_c(Y;X)}{\hat{q}_1(Y;X)}\right)\right] \\
&= -\frac{p}{2}\log c + \frac{(1-c^{-1})}{2\sigma_Y^2}\,\mathbb{E}^{X,Y}\left[\|Y-\delta(X)\|^2\right] \\
&= -\frac{p}{2}\log c + \frac{(1-c^{-1})}{2\sigma_Y^2}\,\mathbb{E}^{X,Y}\left[\|Y-\theta\|^2 + \|\theta-\delta(X)\|^2\right] \qquad (1.13) \\
&= -\frac{p}{2}\log c + \frac{(1-c^{-1})}{2\sigma_Y^2}\left(p\sigma_Y^2 + R(\theta,\delta)\right) \\
&= \frac{p}{2}\,G_{m(\theta)}(c),
\end{aligned}$$

with $m(\theta) = 1 + \frac{R(\theta,\delta)}{p\sigma_Y^2}$. Note that the expectation in (1.13) is obtained using the independence of Y and X given θ and the proof is completed by using the fact that $G_m(c)$ is positive for $1 < c \leq m$, attains its maximum on $(1, \infty)$ at $c = m$, and has a single root on (m, ∞). □

Here are two examples of application of Theorem 1.2.4.

Example 1.2.2. *(Fourdrinier et al. [8]) Suppose model (1.9) with* $p = 1$, $\theta \geq 0$ *and the restricted MLE of* θ, *i.e.* $\delta_{mle}(X) = \max(0, X)$. *A calculation yields*

$$R(\theta, \delta_{mle}) = \theta^2\,\Phi\left(\frac{-\theta}{\sigma_X}\right) + \sigma^2 \int_{-\infty}^{\theta/\sigma} u^2\phi(u)\,du. \qquad (1.14)$$

Since, $\frac{\partial R}{\partial \theta} = 2\theta\,\Phi(\frac{-\theta}{\sigma_X}) > 0$ *the risk function is increasing on* $[0, \infty)$. *Therefore,* $\frac{\sigma_X^2}{2} = R(0, \delta_{mle}) \leq R(\theta, \delta_{mle}) \leq \lim_{\theta\to\infty} R(\theta, \delta_{mle}) = \sigma_X^2$ *and Theorem 1.2.4, tells us,* $\hat{q}_{mle} \sim \mathbb{N}(\delta_{mle}(X), \sigma^2)$ *is dominated by*

$\hat{q}_{mle,c} \sim \mathbb{N}(\delta_{mle}(X), c\sigma^2)$ under KL loss for $1 < c \leq 1 + \frac{\sigma_X^2}{2\,\sigma_Y^2}$ and iff $1 < c \leq c_0(1 + \frac{\sigma_X^2}{2\,\sigma_Y^2})$.
It can derived from a similar analysis to the one above that, among estimators $\hat{q}_{mle,c}$, those with $1 + \frac{\sigma_X^2}{2\sigma_Y^2} \leq c \leq 1 + \frac{\sigma_X^2}{\sigma_Y^2}$ form a complete subclass, while the estimators $\hat{q}_{mle,c}$ with $1 < c \leq c_0(1 + \frac{\sigma_X^2}{2\,\sigma_Y^2})$ form a complete subclass among those that dominate $\hat{q}_{mle}$.

Example 1.2.3. *Let $X_i \sim \mathbb{N}(\theta_i, \sigma^2)$; $i = 1, 2$, independent with $\theta_1 \geq \theta_2$ be independent of $Y \sim \mathbb{N}(\theta_1, \sigma^2)$. The restricted MLE of θ_1 based on $X = (X_1, X_2)$ is given by*

$$\delta_{1,mle}(X_1, X_2) = \begin{cases} X_1 & X_1 \geq X_2 \\ \frac{X_1+X_2}{2} & X_1 < X_2. \end{cases} \tag{1.15}$$

Since $R(\theta, \delta_{1,mle}) = \frac{\sigma^2}{2} + \frac{1}{2}\left(2\sigma^2 + (\theta_1 - \theta_2)^2\right)$ (see Example 2.3.1, for details), we have $R(\theta, \delta_{1,mle}) \geq R(\theta, \delta_{1,mle})\Big|_{\theta_1=\theta_2} = \frac{3\,\sigma^2}{2}$. Hence Theorem 1.2.4 shows that $\hat{q}_{mle,c} \sim \mathbb{N}(\delta_{mle}(X), c\sigma^2)$ dominates $\hat{q}_{mle,c} \sim \mathbb{N}(\delta_{mle}(X), \sigma^2)$ under KL loss for $1 < c \leq 1 + \frac{3}{4}\,\sigma^2$ and iff $1 < c \leq c_0(1 + \frac{3}{4}\sigma^2)$.

Here are some of loss functions for point estimation which will be used in this thesis. The first two arise as dual to predictive density estimation for plug–in densities 1.12, where the latter arises in estimating the ratio of densities (Chapter 4).

(a) *Squared error loss (SEL):* $L(\delta, \theta) = \|\delta - \theta\|^2\,,\ \ \theta \in \mathbb{R}^p$

(b) *Reflected normal loss (RNL):* $L_\gamma(\delta, \theta) = 1 - e^{-\|\delta-\theta\|^2/2\gamma}\,,\ \ \theta \in \mathbb{R}^p$, $\gamma > 0$

(c) *Squared log error loss (Balanced loss) function:* $L(\delta, \theta) = (\log \frac{\delta}{\theta})^2\,,\ \ \theta \in \mathbb{R}$.

1.2.2 Duality between point estimation and predictive density estimation

For normal model (1.9), we focus here the role of the plug–in estimator $\hat{\theta}$ within the predictive density estimator $q_{\hat{\theta},c}$ and review some known duality results with point estimation.

Lemma 1.2.5. (Duality between KL (or RKL) and SEL) *For model (1.9), the frequentist risk of the predictive density estimator $q_{\hat{\theta},c}$ of q_θ under both KL and RKL losses, is dual to the frequentist risk of $\hat{\theta}(X)$ for estimating θ under SEL $\|\hat{\theta}-\theta\|^2$. Namely, $q_{\hat{\theta}_{A,c}}$ dominates $q_{\hat{\theta}_{B,c}}$ under KL (or either RKL) loss iff $\hat{\theta}_A(X)$ dominates $\hat{\theta}_B(X)$ under SEL.*

Proof. We refer to Fourdrinier et al. [8] for the case of KL loss. For RKL loss, the result follows as an application of Theorem 2.6.35; which is a general result for exponential families presented in the Appendix of Chapter 2. □

For $\alpha-$divergence losses whenever $\alpha \in (-1,1)$, it is the reflected normal loss which is dual, as shown by Ghosh, Mergel and Datta [6], as well as Marchand, Perron and Yadegari (2017) for plug-in predictive density estimators, and as expanded upon here for scale expansions in (1.12).

Lemma 1.2.6. (Duality between $\alpha-$divergence and reflected normal losses) *For model (1.1), the frequentist risk of the predictive density estimator $q_{\hat{\theta},c}$ of the density of Y under $\alpha-$divergence loss (1.5), with $\alpha \in (-1,1)$, is dual to the frequentist risk of $\hat{\theta}(X)$ for estimating θ under reflected normal loss $L_{\gamma_0}(\delta,\theta) = 1 - e^{-\|\delta-\theta\|^2/2\gamma_0}$ with $\gamma_0 = (\frac{c^2}{1+\alpha} + \frac{1}{1-\alpha})\,\sigma_Y^2$. Namely, $q_{\hat{\theta}_{A,c}}$ dominates $q_{\hat{\theta}_{B,c}}$ under loss L_α iff $\hat{\theta}_A(X)$ dominates $\hat{\theta}_B(X)$ under loss L_{γ_0} as above.*

Proof. From (1.5), we obtain that the α-divergence losses incurred by the predictive density estimate $q_{\hat{\theta}_1,c^2}$ of $\mathbb{N}_p(\theta,\sigma_Y^2 I_p)$ density is equal to

$$\frac{4}{1-\alpha^2}\left(1-\int_{\mathbb{R}^d}\left(\frac{\phi(\frac{t-\hat{\theta}}{c\sigma_Y})}{(c\sigma_Y)^p}\right)^{\beta}\left(\frac{\phi(\frac{t-\theta}{\sigma_Y})}{\sigma_Y^p}\right)^{1-\beta}dt\right), \tag{1.16}$$

where we have set $\beta=\frac{1+\alpha}{2}$. With the identity $\phi^k(z)=(2\pi)^{\frac{d}{2}(1-k)}\,\phi(z\sqrt{k})$, $k>0$, loss function 1.16 can be written

$$\frac{1}{\beta(1-\beta)}\left(1-(2\pi)^{p/2}\,\frac{\int_{\mathbb{R}^p}\phi(\frac{(t-\hat{\theta})\sqrt{\beta}}{c\sigma_Y})\,\phi(\frac{(t-\theta)\sqrt{1-\beta}}{\sigma_Y})\,dt}{(c^{\beta}\sigma_Y)^d}\right). \tag{1.17}$$

Finally, using the following identity

$$\int_{\mathbb{R}^p}\phi(\frac{t-a_1}{b_1})\,\phi(\frac{t-a_2}{b_2})\,dt=(\frac{b_1^2\,b_2^2}{b_1^2+b_2^2})^{p/2}\,\phi(\frac{a_1-a_2}{\sqrt{b_1^2+b_2^2}}),\,a_i\in\mathbb{R}^p,b_i\in\mathbb{R}_+,i=1,2,$$

for $a_1=\hat{\theta}$, $a_2=\theta$, $b_1=\frac{c\sigma_Y}{\sqrt{\beta}}$, $b_2=\frac{\sigma_Y}{\sqrt{1-\beta}}$, loss function (1.17) reduces to $1-f+fL_{\gamma_0}(\theta,\hat{\theta})$ with $f=\frac{c^{p(1-\beta)}}{(c^2(1-\beta)+\beta)^p}$ and $\gamma_0=\frac{c^2(1-\beta)+\beta}{\beta(1-\beta)}\sigma_Y^2$, thus establishing the result. □

1.3 Point estimation in restricted parameter spaces

Here are some observations on point estimation problems for constrained on parameters which arise in this thesis. Consider a random vector X having a normal distribution $\mathbb{N}_p(\theta,\sigma^2\,I_p)$. In many practical situations, $\theta=(\theta_1,\ldots,\theta_p)$ is restricted to a strict subset of $\mathbb{R}^p$. Some of the common constraints in the literature on the parameter space are as follows:

(i) complete order constraints $\theta_1\geq\theta_2\geq\cdots\geq\theta_p$.

(ii) spherical constraints $\|\theta\| \leq m$, where $\| \cdot \|$ is the Euclidean norm.

(iii) complete (incomplete) order and bounded constraints, $m_1 \geq \theta_1 \geq \theta_2 \geq \cdots \geq \theta_p \geq m_2$.

(iv) umbrella constraints such as, e.g., $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_i \geq \theta_{i+1} \geq \cdots \geq \theta_p$ and for some $i = 2, \ldots, p-1$.

(v) tree order constraints $\theta_1 \leq \theta_i$, for $i = 2, \ldots, p$.

Notice that the benchmark estimator X becomes undesirable and inadmissible under such restrictions. It remains minimax for many cases (see Marchand and Strawderman [14]). This thesis makes use of findings in the restricted parameter literature and useful references include van Eeden [15], as well as Marchand and Strawderman [16]. In particular, we will make use of findings on point estimation with additional information as expanded upon below in Lemma 1.3.8 and 1.3.9, as well as a lovely result due to Hartigan which is as follows.

Theorem 1.3.7 (Hartigan's Theorem). *Let* $X \sim \mathbb{N}_p(\theta, \sigma^2 I_p)$ *and* $\theta \in C$*, where* C *is a convex subset in* $\mathbb{R}^p$ *with non empty interior. Then the Bayes estimator* $\delta(X)$ *with respect to uniform prior on* C *dominates* X *under SEL.*

The following lemma proposes a class of estimators for normal populations which can be transformed to capitalize on estimation problems in constrained parameter spaces. The key technical aspect consists of subdividing the estimation problem into distinct pieces that can be handled separately. This relates to the early work of Blumenthal and Cohen [17], as well as Cohen and Sackrowitz [18]. More recent contributions, using the so–called rotation technique are due to van Eeden and Zidek [19]. Lemma 1.3.8 can be found in Marchand and Strawderman [16].

Lemma 1.3.8. *Suppose* X_1, X_2 *are independently distributed as* $\mathbb{N}_p(\theta_1, \sigma_1^2\, I_p)$ *and* $\mathbb{N}_p(\theta_2, \sigma_2^2\, I_p)$ *respectively with* $\theta_1 - \theta_2 \in A$*, where* A *being a proper subset of* $\mathbb{R}^p$*, and* σ_1^2 *and* σ_2^2 *are known. For estimating* θ_1 *under SEL consider the subclass of estimators*

$$\mathrm{C} = \{\delta_\psi : \delta_\psi(X_1, X_2) = X_2' + \psi(X_1')\}\,, \tag{1.18}$$

where $X_1' = \frac{X_1 - X_2}{1+r}$, $X_2' = \frac{r\,X_1 + X_2}{1+r}$ *and* $r = \frac{\sigma_2^2}{\sigma_1^2}$*. Then* δ_{ψ_1} *dominates* δ_{ψ_0} *in estimating* θ_1*, iff* $\psi_1(X_1')$ *dominates* $\psi_0(X_1')$ *as an estimator of* $\theta_1' = \frac{\theta_1 - \theta_2}{1+r}$ *under the model* $X_1' \sim \mathbb{N}_p(\theta_1' = \frac{\theta_1 - \theta_2}{1+r}, \frac{\sigma_1^2}{1+r}\, I_p)$ *with* $\theta_1' \in \{t : (1+r)t \in A\}$*.*

Proof. It can be seen that X_1' and X_2' are independent and distributed as $\mathbb{N}_p(\theta_1' = \frac{\theta_1 - \theta_2}{1+r}, \frac{\sigma_1^2}{1+r}\, I_p)$ and $\mathbb{N}_p(\theta_2' = \frac{r\theta_1 + \theta_2}{1+r}, \frac{r\sigma_1^2}{1+r}\, I_p)$ respectively. The risk function of δ_ψ can be decomposed as

$$\begin{aligned} R(\delta_\psi, \theta) &= \mathbb{E}_\theta \left(\left\| \left(X_2' - \frac{r\theta_1 + \theta_2}{1+r} \right) + \left(\psi(X_1') - \frac{r\theta_1 - \theta_2}{1+r} \right) \right\|^2 \right) \\ &= \mathbb{E}_\theta \left(\|(X_2' - \theta_2')\|^2 \right) + \mathbb{E}_\theta \left(\|(\psi(X_1') - \theta_1')\|^2 \right), \end{aligned}$$

given the independence of X_1' and X_2', establishing the result. □

As an example, consider $\delta_{\psi_0}(X_1, X_2) = X_1$, i.e. the unrestricted MLE, and A be a convex set with a non empty interior. This estimator belongs to the subclass C in (1.18) with $\psi_0(X_1') = X_1'$. Theorem 1.3.7 applies to $\psi_0(X_1')$ and tells us that the Bayes estimator $\psi_u(X_1')$ of θ_1' with respect to a uniform prior on $\theta_1 - \theta_2 \in A$ dominates

$\psi_0(X_1') = X_1'$. Making use of Lemma 1.3.8 with $\psi_1 = \psi_u$, we obtain

$$\delta_{\psi_u}(X_1, X_2) = X_2' + \psi_u(X_1') \,, \qquad (1.19)$$

which is also Bayes for the problem with prior $\pi(\theta_1, \theta_2) = I_A(\theta_1 - \theta_2)$. For further applications we refer to Marchand and Strawderman [16]. The following lemma relates to the unknown variance σ^2 and represents a novel extension of Marchand et al. [20] to the multivariate case.

Lemma 1.3.9. *Let $X_i \sim \mathbb{N}_p(\theta_i, \sigma^2 I_p)$, $S_i^2 \sim Gamma(\frac{n-1}{2}, 2\sigma^2)$, $i = 1, 2$, $n \geq 2$, be independent with unknown σ^2. Assume that $\theta_1 - \theta_2 \in A$ and the objective is to estimate θ_1 under the loss $\|\frac{\delta - \theta_1}{\sigma}\|^2$. Set,*

$$U_1 = \frac{X_1 - X_2}{2}, \; U_2 = \frac{X_1 + X_2}{2}, \; W = \frac{S_1^2 + S_2^2}{2}, \; \mu_1 = \frac{\theta_1 - \theta_2}{2}, \; \mu_2 = \frac{\theta_1 + \theta_2}{2}.$$

Then, estimators of the form $\delta_\phi(U_1, U_2, W) = U_2 + \phi(U_1, W)$, have risk given by

$$R((\theta_1, \theta_2, \sigma), \delta_\phi) = \frac{1}{2} + \frac{1}{\sigma^2}\, \mathbb{E}^{U_1, W} \|\phi(U_1, W) - \mu_1\|^2.$$

In addition, δ_{ϕ_1} dominates δ_{ϕ_2} iff ϕ_1 dominates ϕ_2 as an estimator of μ_1 under the restriction of $2\mu_1 \in A$, loss $\|\phi - \mu_1\|^2$ based on (U_1, W) with $U_1 \sim \mathbb{N}_p(\mu_1, \frac{\sigma^2}{2})$, $W \sim Gamma(n-1, \sigma^2)$ independent.

Proof. Since $U_i \sim \mathbb{N}(\mu_i, \frac{\sigma^2}{2})$; $i = 1, 2$, and $W \sim Gamma(n-1, \sigma^2)$, are independent, we can write

$$\begin{aligned} \sigma^2 R((\theta_1, \theta_2, \sigma), \delta) &= \mathbb{E}\left[\|U_2 + \phi(U_1, W) - \theta_1\|^2\right] \\ &= \mathbb{E}\left[\|U_2 - \mu_2\|^2 + \|\phi(U_1, W) - \mu_1\|^2\right] \\ &= \frac{p\sigma^2}{2} + \mathbb{E}\left[\|\phi(U_1, W) - \mu_1\|^2\right]. \end{aligned}$$

Furthermore, the dominance result is a direct consequence of the representation of the risk of δ_ϕ. □

Example 1.3.4. *Consider X_1, the MRE estimator of θ_1 under the scaled invariant loss $\|\frac{\delta-\theta_1}{\sigma}\|^2$.*
In the presence of second sample and the additional information $\theta_1 \geq \theta_2$, X_1 is dominated by the class of estimators forming $U_2+\phi(U_1, W)$ according to Lemma 1.3.9. For more information on possible forms of $\phi(U_1, W)$, see Kortbi and Marchand [21].

1.4 Bayes estimators and predictive density estimators in exponential families

We discuss here representations of Bayes point estimators and predictive density estimators for exponential family densities of the form

$$X|\eta \sim p_\eta(x) = h_1(x) \exp\left\{\eta^T s_1(x) - c_1(\eta)\right\} \ , x \in \mathbb{R}^p \, ,$$
$$Y|\eta \sim q_\eta(y) = h_2(y) \exp\{\eta^T s_2(y) - c_2(\eta)\} \, , \ y \in \mathbb{R}^p \, , \tag{1.20}$$

where $\eta \in \mathbb{R}^p$ is a natural parameter. The results of Chapter 4 relate to such densities. Although the predictive densities of Chapters 2 and 3 related to normal models, it is of interest to expand on more general properties for KL and RKL loss functions.
Based on X in model (1.20), there exists an exponential family of conjugate priors

$$\pi(\eta|\,\lambda) = h_0(\eta) \exp\left\{\lambda_1^T \eta + \lambda_2^T(-c(\eta)) - c_0(\lambda)\right\} , \tag{1.21}$$

which also belong to model (1.20), with known $\lambda = (\lambda_1, \lambda_2)^T$. Thus, the posterior densities $\pi(\eta|x, \lambda)$ based on independent $X-$copies

$X_1, \ldots, X_n$, have the form

$$\pi(\eta|x,\lambda) \propto h(\eta)\, \exp\left\{ \lambda_1^{*^T}(x)\eta + \lambda_2^*(-c(\eta)) \right\},$$

where $\lambda_1^*(x) = \lambda_1 + \sum_1^n s_1(x_i)$ and $\lambda_2^* = \lambda_2 + n$. Furthermore, according to (1.2), the Bayesian predictive density for KL loss and prior density π for η with cdf of $G(\cdot)$, is given by

$$\begin{aligned} q(y|x) &= \int p(y|\eta)\, \pi(\eta|x,\lambda)\, dG(\eta) \\ &= \exp\left\{c\left(\lambda_1^*(x) + y,\, \lambda_2^* + 1\right)\right\} / \exp\left\{c\left(\lambda_1^*(x),\, \lambda_2^*\right)\right\}. \end{aligned} \tag{1.22}$$

For more information, see Brown [22].

Next, we consider RKL loss and we are concerned with Bayes predictive density estimators and their frequentist risk. It is known that Bayes predictive density estimators based on model (1.20) are plug–in density estimators as obtained by Yanagimoto and Ohnishi [23].

Theorem 1.4.10. *For model (1.20), the RKL loss, a prior measure π for η such that the posterior exists, the corresponding Bayes predictive density $\hat{q}_\pi$ is the plug-in density $q_{\hat{\eta}}$, with $\hat{\eta}(X) = E_\pi(\eta|X)$ the posterior expectation.*

Proof. See the proof of Theorem 2.6.34. □

Theorem 1.4.11. *For model (1.20), RKL frequentist risk of a plug-in estimator is equivalent to the frequentist risk R_{dual} for the problem of estimating η based on X under the dual loss*

$$L_{dual}(\eta,\hat{\eta}) = \sum_i (\hat{\eta}_i - \eta_i) E_{\hat{\eta}}(s_{2i}(Y)) + (c_2(\eta) - c_2(\hat{\eta}))\,.$$

$q(y\|\eta)$	η	$L_{dual}(\eta, \hat{\eta})$
Poisson(λ)	$\log(\lambda)$	$(\hat{\eta} - \eta)e^{\hat{\eta}} + (e^{\eta} - e^{\hat{\eta}})$
$\mathbb{N}_p(\mu, \nu_Y I_p)$, known ν_Y	μ	$\frac{\\|\hat{\eta}-\eta\\|^2}{2\nu_Y}$
$\mathbb{N}_p(\mu, \nu_Y I_p)$, unknown ν_Y	$\eta_1 = \frac{1}{\nu_Y}; \eta_{i+1} = \frac{\mu_i}{\nu_Y}$	$\frac{p}{2}(\frac{\eta_1}{\hat{\eta}_1} - \log(\frac{\eta_1}{\hat{\eta}_1}) - 1) + \frac{1}{2}\sqrt{\eta_1}\sum_{i\geq 2}(\frac{\eta_i}{\eta_1} - \frac{\hat{\eta}_i}{\hat{\eta}_1})^2$
Binomial(n, p)	$\log(\frac{p}{1-p})$	$n\left((\hat{\eta} - \eta)\frac{e^{\hat{\eta}}}{1+e^{\hat{\eta}}} + \log(\frac{1+e^{\eta}}{1+e^{\hat{\eta}}})\right)$
Gamma(α, β), known α	β	$\alpha\left(\frac{\eta}{\hat{\eta}} - \log(\frac{\eta}{\hat{\eta}}) - 1\right)$
NegativeBinomial(k, p)	$\log(1 - p)$	$k\left\{(\hat{\eta} - \eta)(\frac{e^{\hat{\eta}}}{1-e^{\hat{\eta}}}) + \log(\frac{1-e^{\hat{\eta}}}{1-e^{\eta}})\right\}$
Pareto (pdf $\frac{\alpha}{y^{\alpha+1}}1_{(1,\infty)}(y)$)	α	$\frac{\eta}{\hat{\eta}} - \log(\frac{\eta}{\hat{\eta}}) - 1$

TABLE 1.1: Some densities from model (1.20) with their natural parameters and dual losses.

Furthermore, the Bayes estimator of η under L_{dual} is given by the posterior expectation $E(\eta|X)$.

Proof. See the proof of Theorem 2.6.35. □

Example 1.4.5. *Theorem 1.4.10 tells us that the Bayes density estimator under RKL loss based on model (1.20) is given by the same model but with parameter $\eta(X) = E_\pi(\eta|X)$. For instance, (i) $\hat{q}_\pi(\cdot, X) \sim Poisson(e^{E(\log(\lambda)|X)})$, in the Poisson$(\lambda)$ case, (ii) $\hat{q}_\pi(\cdot, X) \sim \mathbb{N}_p(E(\mu|X)), \nu_Y I_p)$ in the $\mathbb{N}_p(\mu, \nu_Y I_p)$ case with known ν_Y, and (iii) $\hat{q}_\pi(\cdot, X) \sim \mathbb{N}_p(\frac{E(\frac{\mu}{\nu_Y}|X)}{E(\frac{1}{\nu_Y}|X)}, \frac{I_p}{E(\frac{1}{\nu_Y}|X)})$ in the $\mathbb{N}_p(\mu, \nu_Y I_p)$ case with unknown μ, ν_Y. For other examples, we refer to Table 1.1.*

1.5 Skew–normal and other skewed distributions

In this thesis, several skew–normal and skew–Student t distribution will arise, namely generalized Balakrishnan type skew–normal distribution and generalized skew–Student t distribution. These distributions arise below as posterior and predictive distributions. Such relationships have appeared in the large literature on skew–normal distributions (e.g., Liseo and Loperfido [3]).

1.5.1 Generalized Balakrishnan type skew-normal distribution

Definition 1.5.12. A p–variate random variable T is said to have a generalized Balakrishnan type skew–normal distribution, with shape parameters $n \in \mathbb{N}_+$, $\alpha_0 \in \mathbb{R}$, $\alpha_1 \in \mathbb{R}^p$, location and scale parameters $\xi \in \mathbb{R}^p$ and $\tau \in \mathbb{R}$ respectively, denoted by $\mathbb{SN}_p(n, \alpha_0, \alpha_1, \xi, \tau)$, when it admits the pdf

$$\frac{1}{K_n(\alpha_0, \alpha_1)} \frac{1}{\tau^p} \phi_p(\frac{t-\xi}{\tau}) \, \Phi^n(\alpha_0 + \alpha_1^T \frac{t-\xi}{\tau}) \quad t \in \mathbb{R}^p, \tag{1.23}$$

where

$$K_n(\alpha_0, \alpha_1) = \Phi_n\left(\frac{\alpha_0}{\sqrt{1+\alpha_1^T\alpha_1}}, \cdots, \frac{\alpha_0}{\sqrt{1+\alpha_1^T\alpha_1}}; \, \rho = \frac{\alpha_1^T\alpha_1}{1+\alpha_1^T\alpha_1}\right), \tag{1.24}$$

$\phi_p(\cdot)$ is the pdf of a p variate normal distribution, $\Phi(\cdot)$ is cdf of a univariate standard normal density and $\Phi_n(\cdot; \, \rho)$ represents the cdf of a $\mathbb{N}_n(0, \Lambda)$ distribution with covariance matrix $\Lambda = (1-\rho)\, I_n + \rho\, I_n I_n'$.

Remark 1.5.13. *For $\alpha_0 = 0$ and $p = 1$, the densities in (1.23) were proposed by Balakrishnan as a discussant of Arnold and Beaver [24], and further analyzed by Gupta and Gupta [25]. for the normalization constant, we have indeed*

$$\begin{aligned} K_n(\alpha_0, \alpha_1) &= \int_{\mathbb{R}} \phi(z) \Phi^n(\alpha_0 + \alpha_1 z) \, dz \\ &= \mathbb{P}(\cap_{i=1}^n \{U_i \le \alpha_0 + \alpha_1 U_0\}) \\ &= \mathbb{P}(\cap_{i=1}^n \{W_i \le \frac{\alpha_0}{\sqrt{1+\alpha_1^2}}\}) \, , \end{aligned} \tag{1.25}$$

where $(U_0, \ldots, U_n) \sim N_{n+1}(0, I_{n+1})$, $W_i \stackrel{d}{=} \frac{U_i - \alpha_1 U_0}{\sqrt{1+\alpha_1^2}}$, for $i = 1, \ldots, n$, and $(W_1, \ldots, W_n) \sim \mathbb{N}_n(0, \Lambda)$.

Remark 1.5.14. *In Definition 1.5.12, setting* $n = 1$ *yields* $K_1(\alpha_0, \alpha_1) = \Phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^T\alpha_1}})$*. Properties of* $\mathbb{SN}_1(1, \alpha_0, \alpha_1, \xi, \tau)$ *were described by Arnold et al. [26], as well as Arnold and Beaver [24].*

Lemma 1.5.15. *For* $T \sim \mathbb{SN}_p(n, \alpha_0, \alpha_1, \xi, \tau)$ *and* $k \in \mathbb{N}$*, we have*

$$\mathbb{E}(T^k) = \xi + (k-1)\,\mathbb{E}(T^{k-2}) + \frac{\tau n\alpha_1}{\sqrt{1+\alpha_1^2}}\,\phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}})\,\frac{K_{n-1}(\frac{\alpha_0}{1+\alpha_1^2}, \frac{\alpha_1}{\sqrt{1+\alpha_1^2}})}{K_n(\alpha_0, \alpha_1)}\,\mathbb{E}\,(W^{k-1})$$

where $W \sim \mathbb{SN}_p(\frac{n-1,\alpha_0}{1+\alpha_1^2}, \frac{\alpha_0}{1+\alpha_1^2})$*. From this, we have*

$$\mathbb{E}(T) = \xi + \tau\frac{n\alpha_1}{\sqrt{1+\alpha_1^2}}\,\phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}})\,\frac{K_{n-1}(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}, \frac{\alpha_1}{\sqrt{1+\alpha_1^2}})}{K_n(\alpha_0, \alpha_1)}\,. \quad (1.26)$$

Proof. By applying Stein's identity as well as setting $Z = (T-\xi)/\tau$, we have

$$\int z^k\phi(z)\,\Phi^n(\alpha_0+\alpha_1 z)\,dz = \int \phi(z)\frac{\partial}{\partial z}\left(z^{k-1}\Phi^n(\alpha_0+\alpha_1 z)\right)\,dz =$$
$$(k-1)\,\mathbb{E}Z^{k-2}K_n(\alpha_0+\alpha_1) + n\alpha_1\int z^{k-1}\phi(z)\phi(\alpha_0+\alpha_1 z)\,\Phi^{n-1}(\alpha_0+\alpha_1 z)\,dz.$$

The result follows by using the identity $\phi(z)\,\phi(\alpha_0+\alpha_1 z) = \phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}})\,\phi(\sqrt{1+\alpha_1^2}\,z + \frac{\alpha_0\alpha_1}{\sqrt{1+\alpha_1^2}})$, the change of variables $v = \sqrt{1+\alpha_1^2}\,z + \frac{\alpha_0\alpha_1}{\sqrt{1+\alpha_1^2}}$, and the definition of K_{n-1}. □

Remark 1.5.16. *One can verify that setting* $n = 1$ *in equation (1.26), yields*

$$\mathbb{E}(T) = \xi + \tau\frac{\alpha_1}{\sqrt{1+\alpha_1^2}}R(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}})\,, \quad (1.27)$$

where $R(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ *is the inverse Mill's ratio. In addition, the particular case* $\alpha_0 = 0$ *and* $p = 1$ *reduces to the original skew normal density introduced by Azzalini [27].*

Figures 1.1 depicts density (1.23) for $p = 1$, $\alpha_0 = 1, \alpha_1 = 2, \xi = 0$ and $\tau = 1$ for different values of n, while Figures 1.2 and 1.3 illustrate the corresponding density for $n = 3$, $p = 1$ and different values of α_0 and α_1 respectively.

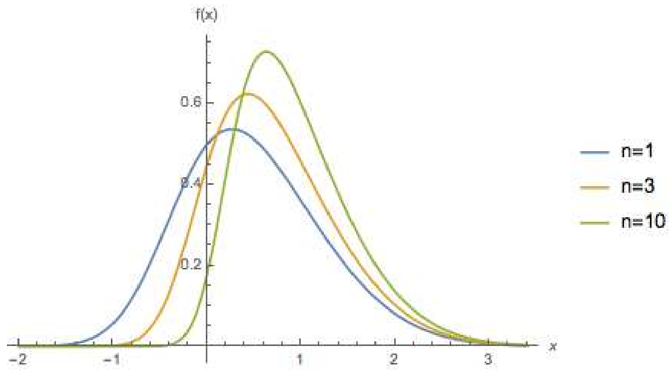

FIGURE 1.1: Density of $\mathbb{SN}_1(n, 1, 2, 0, 1)$ for $n = 1, 3$ and 10.

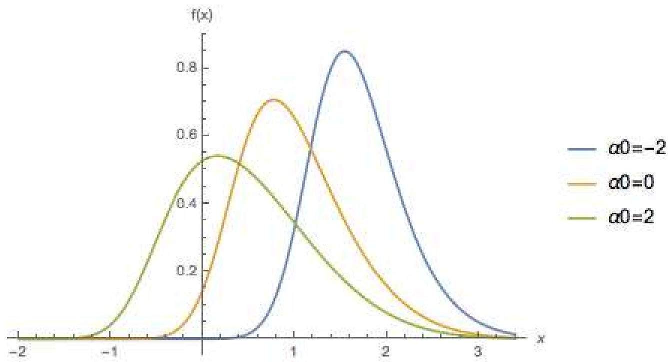

FIGURE 1.2: Density of $\mathbb{SN}_1(3, \alpha_0, 2, 0, 1)$ for $\alpha_0 = -2$, 0 and 2.

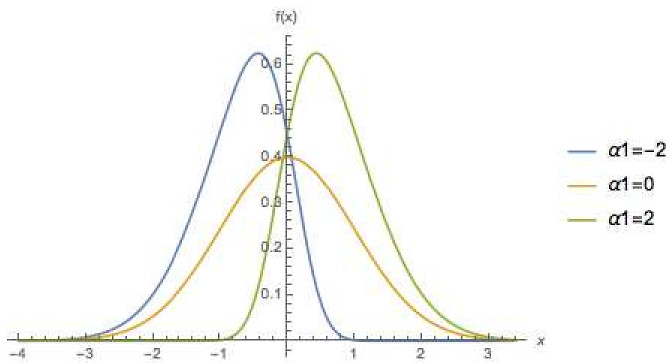

FIGURE 1.3: Density of $\mathbb{SN}_1(3, 1, \alpha_1, 0, 1)$ for $\alpha_1 = -2, 0$ and 2.

There exists other types of skew–normal densities, introduced by Arnold [26], which will arise in Chapter 2, defined as below.

Definition 1.5.17. The density of a skew–normal random variable, with shape parameters $\alpha_0 \in \mathbb{R}, \alpha_1 \in \mathbb{R}^p, \alpha_2 \in \mathbb{R}$, $(\alpha_2 < \alpha_0)$ and location and scale parameters $\xi \in \mathbb{R}^p$ and $\tau \in \mathbb{R}$ respectively, denoted $\mathbb{SN}_p(\alpha_0, \alpha_1, \alpha_2, \xi, \tau)$, is given by

$$\frac{1}{\tau^p}\phi(\frac{t-\xi}{\tau})\,\frac{\Phi(\alpha_0 + \alpha_1^T\frac{t-\xi}{\tau}) - \Phi(\alpha_2 + \alpha_1^T\frac{t-\xi}{\tau})}{\Phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^T\alpha_1}}) - \Phi(\frac{\alpha_2}{\sqrt{1+\alpha_1^T\alpha_1}})}\,,\ t \in \mathbb{R}^p\,. \tag{1.28}$$

As above, one can obtain the expectation

$$\mathbb{E}(T) = \xi + \tau\frac{\alpha_1}{\sqrt{1+\alpha_1^T\alpha_1}}\,\frac{\phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^T\alpha_1}}) - \phi(\frac{\alpha_2}{\sqrt{1+\alpha_1^T\alpha_1}})}{\Phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}) - \Phi(\frac{\alpha_2}{\sqrt{1+\alpha_1^T\alpha_1}})}\,.$$

If we consider $\alpha_2 \to +\infty$, in equation (1.28), we obtain the density in (1.23).

1.5.2 Skew-Student t distribution

Definition 1.5.18 (Student t distribution). A p–variate Student t distribution with degrees of freedom $\nu > 0$, location parameter ξ and scale parameter τ, denoted by $\mathbb{T}_p(\nu, \xi, \tau)$ has density on $\mathbb{R}^p$ given by

$$f_{\nu,\xi,\tau}(t) = \frac{1}{\tau^p}\frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{\frac{d}{2}}}\left(1 + \frac{\|t-\xi\|^2}{\nu\tau^2}\right)^{-\frac{\nu+p}{2}}.$$

Another popular representative of the family of skewed distributions is the skew–Student t distribution, introduced by Azzalini and Capitanio [28], for which the symmetric base distribution is a Student t

distribution.

The following definition introduces general form of skew–Student t distributions related to Arellano-Valle and Genton [29].

Definition 1.5.19 (Skew–Student t distribution). A random variable follows a skew–Student t distribution, denoted by $\mathbb{ST}_p(\nu, \alpha_0, \alpha_1, \xi, \tau)$ with location $\xi \in \mathbb{R}^p$, scale $\tau > 0$, $\nu \in \mathbb{R}_+$, $\alpha_1 \in \mathbb{R}^p$ and $\alpha_0 \in \mathbb{R}$ if $Z = (T - \xi)/\tau$ has pdf

$$f_{\nu,\xi,\tau}(z) \, \frac{F_p\left(\nu + p, \, (\alpha_0 + \alpha_1^T z)\sqrt{\frac{\nu+p}{\nu+z^T z}}\right)}{F_p\left(\nu, \, \frac{\alpha_0}{\sqrt{1+\alpha_1^T \alpha_1}}\right)}, \tag{1.29}$$

where $f_{\nu,\xi,\tau}(\cdot)$ denotes pdf a Student t distribution in Definition 1.5.18 and $F_p(\nu, \cdot)$ is cdf of a standard p–variate Student t distribution. For $p = 1$ and $\alpha_0 = 0$, density (1.29) reduces to

$$2\, t(z, \nu)\, F_1\left(\nu + 1, \, (\frac{\nu + 1}{\nu + z^2})^{\frac{1}{2}} \alpha_1 z\right),$$

(Azzalini and Capitanio [28]), and setting $\alpha_0 = \alpha_1 = 0$ reduces (1.29) to the standard Student t distribution.

We conclude another extensions

Definition 1.5.20. A random variable T follows a skew–Student t distribution, denoted by $\mathbb{ST}_p(\nu, \alpha_0, \alpha_1, \alpha_2, \xi, \tau)$ with location $\xi \in \mathbb{R}^p$, scale $\tau > 0$, $\nu \in \mathbb{R}_+$, $\alpha_1 \in \mathbb{R}^p$, α_0, $\alpha_2 \in \mathbb{R}$, if $Z = (T - \xi)/\tau \in \mathbb{R}^p$, has pdf

$$f_{\nu,\xi,\tau}(z) \, \frac{F_p\left(\nu + p, \, (\alpha_0 + \alpha_1^T z)\sqrt{\frac{\nu+p}{\nu+z^T z}}\right) - F_p\left(\nu + p, \, (\alpha_2 + \alpha_1^T z)\sqrt{\frac{\nu+p}{\nu+z^T z}}\right)}{F_p\left(\nu, \, \frac{\alpha_0}{\sqrt{1+\alpha_1^T \alpha_1}}\right) - F_p\left(\nu, \, \frac{\alpha_2}{\sqrt{1+\alpha_1^T \alpha_1}}\right)}, \tag{1.30}$$

where $f_{\nu,\xi,\tau}(\cdot)$ denotes pdf a Student t distribution in Definition 1.5.18 and $F_p(\nu,\,\cdot)$ are as defined in Definition 1.5.19.

Chapter 2

Predictive Density Estimation When the Variance is Known

This chapter contains a manuscript jointly written with my supervisor Professor Éric Marchand and it includes a substantial number of findings for predictive density estimation of multivariate normal models under α–divergence loss function in the presence of additional information.

On predictive density estimation with additional information [1]

Éric MARCHAND, Abdolnasser SADEGHKHANI
Université de Sherbrooke, Département de mathématiques, Sherbrooke Qc, CANADA, J1K 2R1 (e-mails: eric.marchand@usherbrooke.ca; a.sadeghkhani@usherbrooke.ca)

Summary

Based on independently distributed $X_1 \sim N_p(\theta_1, \sigma_1^2 I_p)$ and $X_2 \sim N_p(\theta_2, \sigma_2^2 I_p)$, we consider the efficiency of various predictive density estimators for $Y_1 \sim N_p(\theta_1, \sigma_Y^2 I_p)$, with the additional information $\theta_1 - \theta_2 \in A$ and known $\sigma_1^2, \sigma_2^2, \sigma_Y^2$. We provide improvements on benchmark predictive densities such as *plug-in*, the maximum likelihood, and the minimum risk equivariant predictive densities. Dominance results are obtained for $\alpha-$divergence losses and include Bayesian improvements for reverse Kullback-Leibler loss, and Kullback-Leibler (KL) loss in the univariate case ($p = 1$). An ensemble of techniques are exploited, including variance expansion (for KL loss), point estimation duality, and concave inequalities. Representations for Bayesian predictive densities, and in particular for $\hat{q}_{\pi_{U,A}}$ associated with a uniform prior for θ truncated to $\{\theta : \theta_1 - \theta_2 \in A\}$, are established and are used for the Bayesian dominance findings. Finally and interestingly, these Bayesian predictive densities also relate to skew-normal distributions, as well as new forms of such distributions.

AMS 2010 subject classifications: 62C20, 62C86, 62F10, 62F15, 62F30

Keywords and phrases: Additional information; α-divergence loss; Bayes estimators; Dominance; Duality; Kullback-Leibler loss; Plug-in; Predictive densities; Restricted parameters; Skew-normal; Variance expansion.

[1] September 2017

2.1 Introduction

2.1.1 Problem and Model

Consider independently distributed

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}_{2p} \left(\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \, \Sigma = \left(\begin{smallmatrix} \sigma_1^2 I_p & 0 \\ 0 & \sigma_2^2 I_p \end{smallmatrix}\right) \right), Y_1 \sim \mathrm{N}_p(\theta_1, \, \sigma_Y^2 I_p) \,, \tag{2.1}$$

where $X_1, X_2, \theta_1, \theta_2$ are $p-$dimensional, and with the additional information (or constraint) $\theta_1 - \theta_2 \in A \subset \mathbb{R}^p$, A, σ_1^2, σ_2^2, σ_Y^2 all known, the variances not necessarily equal. We investigate how to gain from the additional information in providing a predictive density $\hat{q}(\cdot; X)$ as an estimate of the density $q_{\theta_1}(\cdot)$ of Y_1. Such a density is of interest as a surrogate for q_{θ_1}, as well as for generating either future or missing values of Y_1. The additional information $\theta_1 - \theta_2 \in A$ renders X_2 useful in estimating the density of Y_1 despite the independence and the otherwise unrelated parameters.

The reduced X data of the above model is pertinent to summaries X_1 and X_2 that arise through a sufficiency reduction, a large sample approximation, or limit theorems. Specific forms of A include:

(i) order constraints $\theta_{1,i} - \theta_{2,i} \geq 0$ for $i = 1, \ldots, p$; the $\theta_{1,i}$ and $\theta_{2,i}$'s representing the components of θ_1 and θ_2;

(ii) rectangular constraints $|\theta_{1,i} - \theta_{2,i}| \leq m_i$ for $i = 1, \ldots, p$;

(iii) spherical constraints $\|\theta_1 - \theta_2\| \leq m$;

(iv) order and bounded constraints $m_1 \geq \theta_{1,i} \geq \theta_{2,i} \geq m_2$ for $i = 1, \ldots, p$.

There is a very large literature on statistical inference in the presence of such constraints, mostly for **(i)** (e.g., Hwang and Peddada, 1994; Dunson and Neelon, 2003; Park, Kalbfleisch and Taylor, 2014) among many others). Other sources on estimation in restricted parameter spaces can be found in the review paper of Marchand and Strawderman (2004), as well as the monograph by van Eeden (2006). There exist various findings for estimation problems with additional information, dating back to Blumenthal and Cohen (1968) and Cohen and Sackrowitz (1970), with further contributions by van Eeden and Zidek (2001, 2003), Marchand et al. (2008), Marchand and Strawderman (2004).

Remark 2.1.1. *Our set-up applies to various other situations that can be transformed or reduced to model (2.1) with $\theta_1 - \theta_2 \in A$. Here are some examples.*

(I) *Consider model (2.1) with the linear constrained $c_1\theta_1 - c_2\theta_2 + d \in A$, c_1, c_2 being constants not equal to 0, and $d \in \mathbb{R}^p$. Transforming $X_1' = c_1 X_1, X_2' = c_2 X_2 - d$, and $Y_1' = c_1 Y_1$ leads to model (2.1) based on the triplet (X_1', X_2', Y_1'), expectation parameters $\theta_1' = c_1\theta_1, \theta_2' = c_2\theta - d$, covariance matrices $c_i^2\sigma_i^2 I_p, i = 1, 2$ and $c_1^2\sigma_Y^2 I_p$, and with the additional information $\theta_1' - \theta_2' \in A$. With the class of losses being intrinsic (see Remark 2.1.2), and the study of predictive density estimation for Y_1' equivalent to that for Y_1, our basic model and the findings below in this paper will indeed apply for linear constrained $c_1\theta_1 - c_2\theta_2 + d \in A$.*

(II) *Consider a bivariate normal model for X with means θ_1, θ_2, variances σ_1^2, σ_2^2, correlation coefficient $\rho > 0$, and the additional information $\theta_1 - \theta_2 \in A$. The transformation $X_1' = X_1$, $X_2' = \frac{1}{\sqrt{1+\rho^2}}(X_2 - \frac{\rho\sigma_2}{\sigma_1} X_1)$ leads to independent coordinates with means $\theta_1' = \theta_1, \theta_2' = \frac{1}{\sqrt{1+\rho^2}}(\theta_2 - \frac{\rho\sigma_2}{\sigma_1}\theta_1)$, and variances σ_1^2, σ_2^2. We thus obtain model (2.1) with the additional information $\theta_1 - \theta_2 \in A$ transformed to the linear constraint $c_1\theta_1 - c_2\theta_2 + d \in A$, as in part* **(I)** *above, with $c_1 = 1 + \frac{\rho\sigma_2}{\sigma_1}$, $c_2 = \sqrt{1+\rho^2}$, and $d = 0$.*

2.1.2 Predictive density estimation

Several loss functions are at our disposal to measure the efficiency of estimate $\hat{q}(\cdot; x)$, and these include the class of $\alpha-$divergence loss functions (e.g., Csiszàr, 1967) given by

$$L_\alpha(\theta, \hat{q}) = \int_{\mathbb{R}^p} h_\alpha\left(\frac{\hat{q}(y;x)}{q_{\theta_1}(y)}\right) q_{\theta_1}(y)\, dy\,, \tag{2.2}$$

with

$$h_\alpha(z) = \begin{cases} \frac{4}{1-\alpha^2}(1 - z^{(1+\alpha)/2}) & \text{for } |\alpha| < 1 \\ z\log(z) & \text{for } \alpha = 1 \\ -\log(z) & \text{for } \alpha = -1. \end{cases}$$

Notable examples in this class include Kullback-Leibler (h_{-1}), reverse Kullback-Leibler (h_1), and Hellinger ($h_0/4$). For an above given loss, we measure the performance of a predictive density $\hat{q}(\cdot; X)$ by the frequentist risk

$$R_\alpha(\theta, \hat{q}) = \int_{\mathbb{R}^{2p}} L_\alpha\left(\theta, \hat{q}(\cdot; x)\right) \, p_\theta(x)\, dx\,, \tag{2.3}$$

p_θ representing the density of X.

Such a predictive density estimation framework was outlined for Kullback-Leibler loss in the pioneering work of Aitchison and Dunsmore (1975), as well as Aitchison (1975), and has found its way in many different fields of statistical science such as decision theory, information theory, econometrics, machine learning, image processing, and mathematical finance. There has been much recent Bayesian and decision theory analysis of predictive density estimators, in particular for multivariate normal or spherically symmetric settings, as witnessed by the work of Komaki (2001), George, Liang and Xu (2006), Brown, George and Xu (2008), Kato (2009), Fourdrinier et al. (2011), Ghosh, Mergel and Datta (2008), Maruyama and Strawderman (2012), Kubokawa, Marchand and Strawderman (2015, 2017), among others.

Remark 2.1.2. *We point out that losses in (2.2) are intrinsic in the sense that predictive density estimates of the density of $Y' = g(Y)$, with invertible $g : \mathbb{R}^p \to \mathbb{R}^p$ and inverse jacobian J, lead to an equivalent loss with the natural choice $\hat{q}(g^{-1}(y'); x)\,|J|$ as*

$$\int_{\mathbb{R}^p} h_\alpha\left(\frac{\hat{q}(g^{-1}(y'); x)\,|J|}{q_{\theta_1}(g^{-1}(y'))\,|J|}\right) q_{\theta_1}(g^{-1}(y'))\,|J|\, dy' = \int_{\mathbb{R}^p} h_\alpha\left(\frac{\hat{q}(y; x)}{q_{\theta_1}(y)}\right) q_{\theta_1}(y)\, dy\,,$$

which is indeed $L_\alpha(\theta, \hat{q})$ independently of g.

2.1.3 Description of main findings

In our predictive density estimation framework, we study various predictive densities such as: **(i)** *plug-in* densities $N_p(\hat{\theta}_1(X), \sigma_Y^2 I_p)$ including the predictive maximum likelihood estimator (MLE); **(ii)** minimum risk equivariant (MRE) predictive densities $\hat{q}_{mre}$; **(iii)** variance expansions $N_p(\hat{\theta}_1(X), c\sigma_Y^2 I_p)$, with $c > 1$, of *plug-in* predictive densities; and **(iv)** Bayesian predictive densities with an emphasis on the uniform prior for θ truncated to the information set A. Our findings concern, except for Section 2, frequentist risk performance as in (2.3), and related dominated dominance results covering the class of $\alpha-$divergence losses L_α, as well as various types of information sets A.

Subsection 3.1 provides Kullback-Leibler improvements on *plug-in* densities by variance expansion. We make use of a technique due to Fourdrinier et al. (2011), which is universal with respect to p and A and requiring a determination, or lower-bound, of the infimum mean squared error of the plug-in estimator. Such a determination is facilitated by a mean squared error decomposition (Lemma 2.3.14) expressing the risk in terms of the risk of a one-population restricted parameter space estimation problem. Such a decomposition appears in Marchand and Strawderman (2004).

The dominance results of Subsection 3.2 apply to L_α losses and exploit point estimation duality. The targeted predictive densities to be improved upon include *plug-in* densities, $\hat{q}_{mre}$, and more generally predictive densities of the form $\hat{q}_{\hat{\theta}_1,c} \sim N_p(\hat{\theta}_1(X), c\sigma_Y^2 I_p)$. The focus here is on improving on *plug-in* estimates $\hat{\theta}_1(X)$ by exploiting a correspondence with the problem of estimating θ_1 under a dual loss. Both Kullback-Leibler and reverse Kullback-Leibler losses lead to dual mean squared error performance. In turn, as in Marchand and Strawderman (2004), the above risk decomposition relates this performance to a restricted parameter space problem. Results for such problems are thus borrowable to infer dominance results for the original predictive density estimation problem. For other $\alpha-$divergence losses, the strategy is similar, with the added difficulty that the dual loss relates to a reflected normal loss. But, this is handled through a concave inequality technique (e.g., Kubokawa, Marchand and Strawderman, 2015) relating risk comparisons to mean squared error comparisons. Several examples complement the presentation of Section 3.

Sections 2, 4, and 5 relate to Bayesian predictive densities, and especially to the Bayes procedure $\hat{q}_{\pi_{U,A}}$ with respect to the uniform prior $\mathbb{I}_A(\theta_1 - \theta_2)$ restricted to A. Section 2 presents various representations for $\hat{q}_{\pi_{U,A}}$, with examples connecting not only to known skewed-normal distributions, but also to seemingly new families of skewed-normal type distributions. Section 4 contains Bayesian dominance results for both reverse Kullback-Leibler and Kullback-Leibler losses. The case of reverse Kullback-Leibler loss, which is addressed in Subsection 4.1, is special as Bayes predictive densities are necessarily *plug-in* predictive densities, as expanded upon for exponential families in the Appendix. This represents a slight extension of a result due to Yanigimoto and Ohnishi (2009). Moreover, the duality with squared error loss opens the way for Bayesian dominance results. Kullback-Leibler analysis is more challenging, but two dominance findings are obtained in Subsection 4.2.

For $p = 1$, and both $\theta_1 \geq \theta_2$ or $|\theta_1 - \theta_2| \leq m$, using Section 2's representations, we show that the Bayes predictive density $\hat{q}_{\pi_{U,A}}$ improves on $\hat{q}_{mre}$ under Kullback-Leibler loss. Finally, numerical illustrations are presented and commented upon in Section 5.

2.2 Bayesian predictive density estimators and skewed normal type distributions

2.2.1 Bayesian predictive density estimators

We provide here a general representation of the Bayes predictive density estimator of the density of Y_1 in model (2.1) associated with a uniform prior on the additional information set A. Multivariate normal priors truncated to A are plausible choices that are also conjugate, lead to similar results, but will not be further considered in this manuscript. Throughout this manuscript, starting with the next result, we denote ϕ as the $N_p(0, I_p)$ p.d.f.

Lemma 2.2.3. *Consider model (2.1), a Bayes predictive density $\hat{q}_\pi$ with respect to prior π for θ, and the Bayes predictive density $\hat{q}_{\pi_{U,A}}$ with respect to the (uniform) prior $\pi_{U,A}(\theta) = \mathbb{I}_A(\theta_1 - \theta_2)$ for α-divergence loss L_α in (2.2).*

(a) *For $-1 \leq \alpha < 1$, we have*

$$\hat{q}_{\pi_{U,A}}(y_1; x) \propto \hat{q}_{mre}(y_1; x_1)\, I^{\frac{2}{1-\alpha}}(y_1; x)\,, \tag{2.4}$$

with $\hat{q}_{mre}(y_1; x_1)$ the minimum risk predictive density estimator based on x_1 given by a $N_p(x_1, (\sigma_1^2 \frac{(1-\alpha)}{2} + \sigma_Y^2) I_p)$ density, and $I(y_1; x) = \mathbb{P}(T \in A)$, with $T \sim N_p\left(\mu_T, \sigma_T^2 I_p\right)$, $\mu_T = \beta(y_1 - x_1) + (x_1 - x_2)$, $\sigma_T^2 = \frac{2\sigma_1^2 \sigma_Y^2}{(1-\alpha)\sigma_1^2 + 2\sigma_Y^2} + \sigma_2^2$, and $\beta = \frac{(1-\alpha)\sigma_1^2}{(1-\alpha)\sigma_1^2 + 2\sigma_Y^2}$.

(b) *For $\alpha = 1$ (i.e., reverse Kullback-Leibler loss), we have*

$$\hat{q}_\pi(y_1; x) \sim N_p(\mathbb{E}(\theta_1 | x), \sigma_Y^2 I_p)\,, \tag{2.5}$$

where $\mathbb{E}(\theta_1 | x)$ is the posterior expectation of θ_1.

Proof. **(a)** As shown by Corcuera and Giummolè (1999), the Bayes predictive density estimator of the density of Y_1 in (2.1) under loss L_α, $\alpha \neq 1$, is given by

$$\hat{q}_{\pi_{U,A}}(y_1;x) \propto \left\{ \int_{\mathbb{R}^p} \int_{\mathbb{R}^p} \phi^{(1-\alpha)/2}(\frac{y_1-\theta_1}{\sigma_Y})\, \pi(\theta_1,\theta_2|x)\, d\theta_1\, d\theta_2 \right\}^{2/1-\alpha} .$$

With prior measure $\pi_{U,A}(\theta) = \mathbb{I}_A(\theta_1 - \theta_2)$, we obtain

$$\hat{q}_{\pi_{U,A}}(y_1;x) \propto \left\{ \int_{\mathbb{R}^p} \int_{\mathbb{R}^p} \phi(\frac{y_1-\theta_1}{\sqrt{\frac{2}{1-\alpha}\sigma_Y^2}})\, \phi(\frac{\theta_1-x_1}{\sigma_1})\, \phi(\frac{\theta_2-x_2}{\sigma_2})\, \mathbb{I}_A(\theta_1-\theta_2)\; d\theta_1\, d\theta_2 \right\}^{2/1-\alpha} ,$$

given that $\phi^m(z) \propto \phi(m^{1/2}z)$ for $m > 0$. By the decomposition

$$\frac{\|\theta_1 - y_1\|^2}{a} + \frac{\|\theta_1 - x_1\|^2}{b} = \frac{\|y_1 - x_1\|^2}{a+b} + \frac{\|\theta_1 - w\|^2}{\sigma_w^2},$$

with $a = \frac{2\sigma_Y^2}{1-\alpha}$, $b = \sigma_1^2$, and $w = \frac{by_1+ax_1}{a+b} = \beta y_1 + (1-\beta)x_1$, $\sigma_w^2 = \frac{ab}{a+b} = \frac{2\sigma_1^2\sigma_Y^2}{2\sigma_Y^2+(1-\alpha)\sigma_1^2}$, we obtain

$$\begin{aligned} \hat{q}_{\pi_{U,A}}(y_1;x) &\propto \phi^{2/(1-\alpha)}(\frac{y_1-x_1}{\sqrt{\frac{2\sigma_Y^2}{1-\alpha}+\sigma_1^2}}) \left\{ \int_{\mathbb{R}^{2p}} \phi(\frac{\theta_1-w}{\sigma_w})\, \phi(\frac{\theta_2-x_2}{\sigma_2})\, \mathbb{I}_A(\theta_1-\theta_2)\; d\theta_1\, d\theta_2 \right\}^{2/1-\alpha} \\ &\propto \hat{q}_{\mathrm{mre}}(y_1;x_1)\, \{\mathbb{P}(Z_1 - Z_2 \in A)\}^{2/1-\alpha} , \end{aligned}$$

with Z_1, Z_2 independently distributed as $Z_1 \sim N_p(w, \sigma_w^2)$, $Z_2 \sim N_p(x_2, \sigma_2^2)$. The result follows by setting $T =^d Z_1 - Z_2$.

(b) This part is a consequence of Theorem 2.6.34, which is a general result for exponential families; presented in the Appendix; and which establishes that Bayes predictive densities are necessarily *plug–in* predictive densities. See Example 2.6.5 for details. □

The general form of the Bayes predictive density estimator $\hat{q}_{\pi_{U,A}}$ is thus a weighted version of $\hat{q}_{\mathrm{mre}}$, with the weight a multivariate normal probability raised to the $2/(1-\alpha)^{th}$ power which is a function of y_1 and which depends on x, α, A. Observe that the representation applies in the trivial case $A = \mathbb{R}^p$, yielding $I = 1$ and $\hat{q}_{\mathrm{mre}}$ as the Bayes estimator. As expanded on in Subsection 2.2.2, the densities $\hat{q}_{\pi_{U,A}}$ for Kullback-Leibler loss relate to skew-normal distributions, and more generally to skewed distributions arising from selection (see for instance Arnold and Beaver, 2002; Arellano-Valle, Branco and Genton, 2006; among others). Moreover, it is known (e.g. Liseo and Loperfido, 2003) that posterior distributions present here also relate to such skew-normal type distributions. Lemma 2.2.3 does not address the evaluation of the normalization constant for the Bayes predictive density $\hat{q}_{\pi_{U,A}}$, but we now proceed with this for the particular cases of Kullback-Leibler and Hellinger losses, and more generally for cases where $\frac{2}{1-\alpha}$ is a positive integer, i.e., $\alpha = 1 - \frac{2}{n}$ where $n = 1, 2, \ldots$. In what follows, we denote 1_m as the m dimensional column vector with components equal to 1, and $\otimes$ as the usual Kronecker product.

Lemma 2.2.4. *For model (2.1), $\alpha-$divergence loss with $n = \frac{2}{1-\alpha} \in \{1, 2, \ldots\}$, the Bayes predictive density $\hat{q}_{\pi_{U,A}}(y_1; x)\,, y_1 \in \mathbb{R}^p$, with respect to the (uniform) prior $\pi_{U,A}(\theta) = \mathbb{I}_A(\theta_1 - \theta_2)$, is given by*

$$\hat{q}_{\pi_{U,A}}(y_1; x) \;=\; \hat{q}mre(y_1; x_1)\, \frac{\{\mathbb{P}(T \in A)\}^n}{\mathbb{P}(\cap_{i=1}^n \{Z_i \in A\})}\,, \tag{2.6}$$

with $\hat{q}\mathrm{mre}(y_1; x_1)$ a $N_p(x_1, (\sigma_1^2/n + \sigma_Y^2)\mathrm{I}_p)$ density, $T \sim N_p(\mu_T, \sigma_T^2 I_p)$ with $\mu_T = \beta(y_1 - x_1) + (x_1 - x_2)$, $\sigma_T^2 = \sigma_2^2 + n\sigma_Y^2\beta$, $\beta = \frac{\sigma_1^2}{\sigma_1^2 + n\sigma_Y^2}$, and $Z = (Z_1, \ldots, Z_n)' \sim N_{np}(\mu_Z, \Sigma_Z)$ with $\mu_Z = 1_n \otimes (x_1 - x_2)$ and $\Sigma_Z = (\sigma_T^2 + \sigma_Y^2\beta^2) I_{np} + (\frac{\beta^2\sigma_1^2}{n} 1_n 1_n' \otimes \mathrm{I}_p)$

Remark 2.2.5. *The Kullback-Leibler case corresponds to $n = 1$ and the above form of the Bayes predictive density simplifies to*

$$\hat{q}_{\pi_{U,A}}(y_1; x) \;=\; \hat{q}mre(y_1; x_1)\, \frac{\mathbb{P}(T \in A)}{\mathbb{P}(Z_1 \in A)}\,, \tag{2.7}$$

with $\hat{q}\mathrm{mre}(y_1; x_1)$ a $N_p(x_1, (\sigma_1^2 + \sigma_Y^2)\mathrm{I}_p)$ density, $T \sim N_p(\mu_T, \sigma_T^2 I_p)$ with $\mu_T = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_Y^2}(y_1 - x_1) + (x_1 - x_2)$ and $\sigma_T^2 = \frac{\sigma_1^2\sigma_Y^2}{\sigma_1^2 + \sigma_Y^2} + \sigma_2^2$, and $Z_1 \sim N_p(x_1 - x_2, (\sigma_1^2 + \sigma_2^2) I_p)$. In the univariate case (i.e., $p = 1$), T is univariate normally distributed and the expectation and covariance matrix of Z simplify to $1_n(x_1 - x_2)$ and $(\sigma_T^2 + \sigma_Y^2\beta^2) I_n + \beta^2 \frac{\sigma_1^2}{n} 1_n 1_n'$ respectively. Finally, we point out that the diagonal elements of Σ_Z simplify to $\sigma_1^2 + \sigma_2^2$, a result which will arise below several times.

Proof of Lemma 2.2.4. It suffices to evaluate the normalization constant (say C) for the predictive density in (2.4). We have

$$\begin{aligned} C &= \int_{\mathbb{R}^p} \hat{q}\mathrm{mre}(y_1; x_1)\, \{\mathbb{P}(T \in A)\}^n\, dy_1 \\ &= \int_{\mathbb{R}^p} \hat{q}\mathrm{mre}(y_1; x_1)\, \mathbb{P}\left(\cap_{i=1}^n \{T_i \in A\}\right)\, dy_1\,, \end{aligned}$$

with $T_1, \ldots, T_n$ independent copies of T. With the change of variables $u_0 = \frac{y_1 - x_1}{\sqrt{\sigma_1^2/n + \sigma_Y^2}}$ and letting $U_0, U_1, \ldots, U_n$ i.i.d. $N_p(0, I_p)$, we obtain

$$
\begin{aligned}
C &= \int_{\mathbb{R}^p} \phi(u_0)\, \mathbb{P}\left(\cap_{i=1}^n \{\sigma_T U_i + \beta u_0 \sqrt{\sigma_1^2/n + \sigma_Y^2} + x_1 - x_2\} \in A\right) du_0 \\
&= \mathbb{P}\left(\cap_{i=1}^n \{\sigma_T U_i + \beta U_0 \sqrt{\sigma_1^2/n + \sigma_Y^2} + x_1 - x_2\} \in A\right), \\
&= \mathbb{P}\left(\cap_{i=1}^n \{Z_i \in A\}\right).
\end{aligned}
$$

The result follows by verifying that the expectation and covariance matrix of $Z = (Z_1, \ldots, Z_n)'$ are as stated. □

The next result presents a useful posterior distribution decomposition, with an accompanying representation of the posterior expectation $\mathbb{E}(\theta_1|x)$ in terms of a truncated multivariate normal expectation. The latter characterizes the Bayes predictive density under reverse Kullback-Leibler loss in accordance with Lemma 2.2.3, as well as coincide with the expectation under the Bayes Kullback-Leibler predictive density $\hat{q}_{\pi_{U,A}}$. Specific examples will be presented in Subsection 2.3.4.

Lemma 2.2.6. *Consider $X|\theta$ as in model (2.1) and the uniform prior $\pi_{U,A}(\theta) = \mathbb{I}_A(\theta_1 - \theta_2)$. Set $r = \frac{\sigma_2^2}{\sigma_1^2}$, $\omega_1 = \theta_1 - \theta_2$, and $\omega_2 = r\theta_1 + \theta_2$. Then, conditional on $X = x$, ω_1 and ω_2 are independently distributed with*

$$\omega_1 \sim N_p(\mu_{\omega_1}, \tau_{\omega_1}^2) \text{ truncated to } A, \qquad \omega_2 \sim N_p(\mu_{\omega_2}, \tau_{\omega_2}^2),$$

$\mu_{\omega_1} = x_1 - x_2$, $\mu_{\omega_2} = rx_1 + x_2$, $\tau_{\omega_1}^2 = \sigma_1^2 + \sigma_2^2$, and $\tau_{\omega_2}^2 = 2\sigma_2^2$. Furthermore, we have $\mathbb{E}(\theta_1|x) = \frac{1}{1+r}\left(\mathbb{E}(\omega_1|x) + \mu_{\omega_2}\right)$.

Proof. With the posterior density $\pi(\theta|x) \propto \phi(\frac{\theta_1 - x_1}{\sigma_1})\ \phi(\frac{\theta_2 - x_2}{\sigma_2})\ \mathbb{I}_A(\theta_1 - \theta_2)$, the result follows by transforming to (ω_1, ω_2). □

2.2.2 Examples of Bayesian predictive density estimators

With the presentation of the Bayes predictive estimator $\hat{q}_{\pi_{U,A}}$ in Lemmas 2.2.3 and 2.2.4, which is quite general with respect to the dimension p, the additional information set A, and the $\alpha-$divergence loss, it is pertinent and instructive to continue with some illustrations. Moreover, various skewed-normal or skewed-normal type, including new extensions, arise as predictive density estimators. Such distributions have indeed generated much interest for the last thirty years or so, and continue to do so, as witnessed by the large literature devoted to their study. The most familiar choices of $\alpha-$divergence loss are Kullback-Leibler and Hellinger (i.e., $n = \frac{2}{1-\alpha} = 1, 2$ below) but the form of the Bayes predictive density estimator $\hat{q}_{\pi_{U,A}}$ is nevertheless expanded upon below in the context of Lemma 2.2.4, in view of the connections with an extended family of skewed-normal type distributions (e.g., Definition 2.2.7), which is also of independent interest.

Subsections 2.2.1, 2.2.2, 2.2.3. deal with Kullback-Leibler and $\alpha-$divergence losses for situations: (i) $p = 1, A = \mathbb{R}_+$; (ii) $p = 1, A = [-m, m]$; (iii) $p \geq 1$ and A a ball of radius m centered at the origin, while Subsection 2.2.4. deals with reverse Kullback-Leibler loss.

2.2.2.1 Univariate case with $\theta_1 \geq \theta_2$

From (2.6), we obtain for $p = 1, A = \mathbb{R}_+$: $\mathbb{P}(T \in A) = \Phi(\frac{\mu_T}{\sigma_T})$ and

$$\hat{q}_{\pi_{U,A}}(y_1; x) \propto \frac{1}{\sqrt{\sigma_1^2/n + \sigma_Y^2}}\, \phi(\frac{y_1 - x_1}{\sqrt{\sigma_1^2/n + \sigma_Y^2}})\, \Phi^n(\frac{\beta(y_1 - x_1) + (x_1 - x_2)}{\sigma_T}), \quad (2.8)$$

with β and σ_T^2 given in Lemma 2.2.4. These densities match the following family of densities.

Definition 2.2.7. A generalized Balakrishnan type skewed-normal distribution, with shape parameters $n \in \mathbb{N}_+, \alpha_0, \alpha_1 \in \mathbb{R}$, location and scale parameters ξ and τ, denoted $\mathrm{SN}(n, \alpha_0, \alpha_1, \xi, \tau)$, has density on $\mathbb{R}$ given by

$$\frac{1}{K_n(\alpha_0, \alpha_1)}\, \frac{1}{\tau}\, \phi(\frac{t-\xi}{\tau})\, \Phi^n(\alpha_0 + \alpha_1 \frac{t-\xi}{\tau})\,, \quad (2.9)$$

with

$$K_n(\alpha_0, \alpha_1) = \Phi_n\left(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}, \cdots, \frac{\alpha_0}{\sqrt{1+\alpha_1^2}}; \rho = \frac{\alpha_1^2}{1+\alpha_1^2}\right),$$

$\Phi_n(\cdot; \rho)$ representing the cdf of a $\mathrm{N}_n(0, \Lambda)$ distribution with covariance matrix $\Lambda = (1-\rho)\, I_n + \rho\, 1_n 1_n'$.

Remark 2.2.8. *(The case $n = 1$)*
$\mathrm{SN}(1, \alpha_0, \alpha_1, \xi, \tau)$ *densities are given by (2.9) with* $n = 1$ *and* $K_1(\alpha_0, \alpha_1) = \Phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}})$. *Properties of* $\mathrm{SN}(1, \alpha_0, \alpha_1, \xi, \tau)$ *distributions were described by Arnold et al. (1993), as well as Arnold and Beaver (2002), with the particular case* $\alpha_0 = 0$ *reducing to the original skew normal density modulo a location-scale transformation as presented in Azzalini's seminal 1985 paper. Namely, the expectation of* $T \sim \mathrm{SN}(1, \alpha_0, \alpha_1, \xi, \tau)$ *is given by*

$$\mathbb{E}(T) = \xi + \tau \frac{\alpha_1}{\sqrt{1+\alpha_1^2}} R(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}), \tag{2.10}$$

with $R =: \frac{\phi}{\Phi}$ *known as the inverse Mill's ratio.*

Remark 2.2.9. *For* $\alpha_0 = 0, n = 2, 3, \ldots,$ *the densities were proposed by Balakrishnan as a discussant of Arnold and Beaver (2002), and further analyzed by Gupta and Gupta (2004). We are not aware of an explicit treatment of such distributions in the general case, but standard techniques may be used to derive the following properties. For instance, as handled more generally above in the proof of Lemma 2.2.4, the normalization constant* K_n *may be expressed in terms of a multivariate normal c.d.f. by observing that*

$$\begin{aligned} K_n(\alpha_0, \alpha_1) &= \int_{\mathbb{R}} \phi(z) \Phi^n(\alpha_0 + \alpha_1 z)\, dz \\ &= \mathbb{P}(\cap_{i=1}^n \{U_i \leq \alpha_0 + \alpha_1 U_0\}) \\ &= \mathbb{P}(\cap_{i=1}^n \{W_i \leq \frac{\alpha_0}{\sqrt{1+\alpha_1^2}}\}), \end{aligned}$$

with $(U_0, \ldots, U_n) \sim N_{n+1}(0, I_{n+1})$, $W_i \stackrel{d}{=} \frac{U_i - \alpha_1 U_0}{\sqrt{1+\alpha_1^2}}$, *for* $i = 1, \ldots, n$, *and* $(W_1, \ldots, W_n) \sim \mathrm{N}_n(0, \Lambda)$.

In terms of expectation, we have, for $T \sim SN(n, \alpha_0, \alpha_1, \xi, \tau)$, $\mathbb{E}(T) = \xi + \tau \mathbb{E}(W)$ *where* $W \sim SN(n, \alpha_0, \alpha_1, 0, 1)$ *and*

$$\mathbb{E}(W) = \frac{n\alpha_1}{\sqrt{1+\alpha_1^2}} \phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}) \frac{K_{n-1}(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}, \frac{\alpha_1}{\sqrt{1+\alpha_1^2}})}{K_n(\alpha_0, \alpha_1)}. \tag{2.11}$$

This can be obtained via Stein's identity $\mathbb{E}\,Ug(U) = \mathbb{E}g'(U)$ for differentiable g and $U \sim N(0,1)$. Indeed, we have

$$\int_{\mathbb{R}} u\phi(u)\,\Phi^n(\alpha_0+\alpha_1 u)\,du \;=\; n\alpha_1 \int_{\mathbb{R}} \phi(u)\phi(\alpha_0+\alpha_1 u)\,\Phi^{n-1}(\alpha_0+\alpha_1 u)\,du$$

and the result follows by making use of the identity $\phi(u)\,\phi(\alpha_0+\alpha_1 u) = \phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}})\,\phi(v)$, with $v = \sqrt{1+\alpha_1^2}\,u + \frac{\alpha_0\alpha_1}{\sqrt{1+\alpha_1^2}}$, as well as the change of variables $u \to v$ and the definition of K_{n-1}.

The connection between the densities of Definition 2.2.7 and the predictive densities in (2.8) is thus explicitly stated as follows, with the Kullback-Leibler and Hellinger cases corresponding to $n = 1, 2$ respectively.

Corollary 2.2.10. *For $p = 1$, $A = \mathbb{R}_+$, $\pi_{U,A}(\theta) = \mathbb{I}_A(\theta_1-\theta_2)$, the Bayes predictive density estimator $\hat{q}_{\pi_{U,A}}$ under $\alpha-$divergence loss, with $n = \frac{2}{1-\alpha} \in \mathbb{N}_+$ positive integer, is given by a* $\mathrm{SN}(n, \alpha_0 = \frac{x_1-x_2}{\sigma_T}, \alpha_1 = \frac{\beta\tau}{\sigma_T}, \xi = x_1, \tau = \sqrt{\frac{\sigma_1^2}{n}+\sigma_Y^2})$ *density, with $\sigma_T^2 = \sigma_2^2 + n\beta\sigma_Y^2$ and $\beta = \frac{\sigma_1^2}{\sigma_1^2+n\sigma_Y^2}$.*

Remark 2.2.11. *For the equal variances case with $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2 = \sigma^2$, the above predictive density estimator is a* $\mathrm{SN}(n, \alpha_0 = \sqrt{\frac{n+1}{(2n+1)\sigma}}(x_1-x_2), \alpha_1 = \sqrt{\frac{1}{n(2n+1)}}, \xi = x_1, \tau = \sqrt{\frac{n+1}{n}}\sigma)$ *density.*

2.2.2.2 Univariate case with $|\theta_1 - \theta_2| \leq m$

From (2.6), we obtain for $p = 1, A = [-m, m]$: $\mathbb{P}(T \in A) = \Phi(\frac{\mu_T+m}{\sigma_T}) - \Phi(\frac{\mu_T-m}{\sigma_T})$, and we may write

$$\hat{q}_{\pi_{U,A}}(y_1; x) = \frac{1}{\tau}\phi(\frac{t-\xi}{\tau})\,\frac{\{\Phi(\alpha_0+\alpha_1\frac{t-\xi}{\tau}) - \Phi(\alpha_2+\alpha_1\frac{t-\xi}{\tau})\}^n}{J_n(\alpha_0,\alpha_1,\alpha_2)}\,, \tag{2.12}$$

with $\xi = x_1, \tau = \sqrt{\sigma_1^2/n+\sigma_Y^2}, \alpha_0 = \frac{x_1-x_2+m}{\sigma_T}$, $\alpha_1 = \frac{\beta\tau}{\sigma_T}$ $\alpha_2 = \frac{x_1-x_2-m}{\sigma_T}$, β and σ_T^2 given in Lemma 2.2.4, and $J_n(\alpha_0,\alpha_1,\alpha_2)$ (independent of ξ,τ) a special case of the normalization constant given in (2.6).

For fixed n, the densities in (2.12) form a five-parameter family of densities with location and scale parameters $\xi \in \mathbb{R}$ and $\tau \in \mathbb{R}_+$, and shape parameters $\alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$ such that $\alpha_0 > \alpha_2$. The Kullback-Leibler predictive densities

($n = 1$) match densities introduced by Arnold et al. (1993) with the normalization constant in (2.12) simplifying to:

$$J_1(\alpha_0, \alpha_1, \alpha_2) = \Phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}) - \Phi(\frac{\alpha_2}{\sqrt{1+\alpha_1^2}}) = \Phi(\frac{m-(x_1-x_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}) - \Phi(\frac{-m-(x_1-x_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}). \tag{2.13}$$

The corresponding expectation is readily obtained as in (2.10) and equals

$$\begin{aligned} \mathbb{E}(T) &= \xi + \tau \frac{\alpha_1}{\sqrt{1+\alpha_1^2}} \frac{\phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}) - \phi(\frac{\alpha_2}{\sqrt{1+\alpha_1^2}})}{\Phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}) - \Phi(\frac{\alpha_2}{\sqrt{1+\alpha_1^2}})} \\ &= x_1 + \frac{\sigma_1^2}{\sqrt{\sigma_1^2+\sigma_2^2}} \frac{\phi(\frac{x_1-x_2+m}{\sqrt{\sigma_1^2+\sigma_2^2}}) - \phi(\frac{x_1-x_2-m}{\sqrt{\sigma_1^2+\sigma_2^2}})}{\Phi(\frac{x_1-x_2+m}{\sqrt{\sigma_1^2+\sigma_2^2}}) - \Phi(\frac{x_1-x_2-m}{\sqrt{\sigma_1^2+\sigma_2^2}})}, \end{aligned} \tag{2.14}$$

by using the above values of $\xi, \tau, \alpha_0, \alpha_1, \alpha_2$.

Hellinger loss yields the Bayes predictive density in (2.12) with $n = 2$, and a calculation as in Remark 2.2.9 leads to the evaluation

$$J_2(\alpha_0, \alpha_1, \alpha_2) = \Phi_2(\alpha_0', \alpha_0'; \alpha_1') + \Phi_2(\alpha_2', \alpha_2'; \alpha_1') - 2\Phi_2(\alpha_0', \alpha_2'; \alpha_1')$$

with $\alpha_i' = \frac{\alpha_i}{\sqrt{1+\alpha_1^2}}$ for $i = 0, 1, 2$.

2.2.2.3 Multivariate case with $||\theta_1 - \theta_2|| \leq m$

For $p \geq 1$, the ball $A = \{t \in \mathbb{R}^p : ||t|| \leq m\}$, μ_T and σ_T^2 as given in Lemma 2.12, the Bayes predictive density in (2.6) under $\alpha-$divergence loss with $\frac{2}{1-\alpha} = n \in \mathbb{N}_+$ is expressible as

$$\hat{q}_{\pi_{U,A}} \propto \hat{q}_{\text{mre}}(y_1; x_1) \{\mathbb{P}(||T||^2 \leq m^2)\}^n$$

with $T \sim \sigma_T^2 \chi_p^2(\|\mu_T\|^2/\sigma_T^2)$, i.e., the weight attached to $\hat{q}_{\text{mre}}$ is proportional to the n^{th} power of the c.d.f. of a non-central chi-square distribution.

For Kullback-Leibler loss, we obtain from (2.6)

$$\begin{aligned} \hat{q}_{\pi_{U,A}}(y_1; x) &= \hat{q}_{\text{mre}}(y_1; x_1) \frac{\mathbb{P}(||T||^2 \leq m^2)}{\mathbb{P}(||Z_1||^2 \leq m^2)} \\ &= \hat{q}_{\text{mre}}(y_1; x_1) \frac{\mathbb{F}_{p,\lambda_1(x,y_1)}(m^2/\sigma_T^2)}{\mathbb{F}_{p,\lambda_2(x)}(m^2/(\sigma_1^2+\sigma_2^2))}, \end{aligned} \tag{2.15}$$

where $F_{p,\lambda}$ represents the c.d.f. of a $\chi^2_p(\lambda)$ distribution, $\lambda_1(x,y_1) = \frac{\|\mu_T\|^2}{\sigma_T^2}$ and $\lambda_2(x) = \frac{\|x_1-x_2\|^2}{\sigma_1^2+\sigma_2^2}$. Observe that the non-centrality parameters λ_1 and λ_2 are random, and themselves non-central chi-square distributed as $\lambda_1(X,Y_1) \sim \chi^2_p(\frac{||\theta_1-\theta_2||^2}{\sigma_T^2})$ and $\lambda_2(X) \sim \chi^2_p(\frac{||\theta_1-\theta_2||^2}{\sigma_1^2+\sigma_2^2})$.

Of course, the above predictive density (2.15) matches the Kullback-Leibler predictive density given in (2.12) for $n=1$, and represents an otherwise interesting multivariate extension.

2.2.2.4 REVERSE KULLBACK-LEIBLER LOSS

It follows from Lemma 2.2.3 and Lemma 2.2.6 (also see Lemma 2.4.23) that the Bayes predictive density estimator $\hat{q}_{\pi_{U,A}}$ for reverse Kullback-Leibler loss, is given by a $N_p(\mathbb{E}(\theta_1|x), \sigma_Y^2 I_p)$ density with

$$\mathbb{E}(\theta_1|x) = \frac{1}{1+r}\,(\mathbb{E}(\omega_1|x)+rx_1+x_2), \ \text{ with } \omega_1 \sim N_p(x_1-x_2, (\sigma_1^2+\sigma_2^2)I_p) \text{ truncated to } A\,. \tag{2.16}$$

Truncated normal distributions and their expectations are familiar quantities and thus provide expressions for such predictive densities. Alternatively, as mentioned in the paragraph preceding Lemma 2.2.6, the expectation $\mathbb{E}(\theta_1|x)$ also matches the expected value under the Kullback-Leibler Bayes predictive density $\hat{q}_{U,A}$. We illustrate these two above approaches by evaluating (2.16) for the following situations.

(I) Consider $p=1, A=\mathbb{R}_+$ and let $T \sim \hat{q}_{\pi_{U,A}}$ corresponding to Kullback-Leibler loss. Then, we have

$$\mathbb{E}(\theta_1|x) = \mathbb{E}(T) = x_1 + \frac{\sigma_1^2}{\sqrt{\sigma_1^2+\sigma_2^2}}\,R(\frac{x_1-x_2}{\sqrt{\sigma_1^2+\sigma_2^2}})\,,$$

by using directly (2.10) and Corollary 2.2.10.

(II) Similarly, for $p=1, A=[-m,m]$, letting let $T \sim \hat{q}_{\pi_{U,A}}$ corresponding to Kullback-Leibler loss, we have $\mathbb{E}(\theta_1|x) = \mathbb{E}(T)$ as given in (2.14).

(III) Consider the ball $A = \{t \in \mathbb{R}^p : \|t\| \leq m\}$ with $p \geq 1$. Observe that $\mathbb{E}(\omega_1|x) = \delta_{\pi_{U,A}}(x')$, with $x' = x_1 - x_2$, is the Bayes point estimator under squared error loss based on the model $X' \sim N_p(\mu, (\sigma_1^2+\sigma_2^2)I_p)$ and the prior

$\pi_{U,A}$. Such an estimator was expressed in terms of the $\chi^2_p(\lambda)$ c.d.f. $F_{p,\lambda}$ by Marchand and Perron (2001, Remark 1). From their formula and the above connection, we obtain an evaluation of (2.16) with

$$\mathbb{E}(\omega_1|x) \;=\; (x_1 - x_2)\,\frac{F_{p+2,\frac{\|x_1-x_2\|^2}{\sigma_1^2+\sigma_2^2}}\left(\frac{m^2}{\sigma_1^2+\sigma_2^2}\right)}{F_{p,\frac{\|x_1-x_2\|^2}{\sigma_1^2+\sigma_2^2}}\left(\frac{m^2}{\sigma_1^2+\sigma_2^2}\right)}\,.$$

2.3 General dominance results

We exploit different channels to obtain predictive density estimation improvements on benchmark procedures such as the maximum likelihood predictive density estimator $\hat{q}_{\text{mle}}$ and the minimum risk equivariant predictive density $\hat{q}_{\text{mre}}$. These predictive density estimators are members of the larger class of densities

$$q_{\hat{\theta}_1,c} \sim N_p(\hat{\theta}_1(X), c\sigma_Y^2 I_p)\,, \tag{2.17}$$

with, for instance, the choice $\hat{\theta}_1(X) = \hat{\theta}_{1,\text{mle}}(X), c = 1$ yielding $\hat{q}_{\text{mle}}$, and $\hat{\theta}_1(X) = X, c = 1 + \frac{(1-\alpha)\sigma_1^2}{2\sigma_Y^2}$ yielding $\hat{q}_{\text{mre}}$ for loss L_α. Two main strategies are exploited to produce improvements: **(A)** scale expansion and **(B)** point estimation duality.

(A) *Plug–in* predictive densities $q_{\hat{\theta}_1,1}$ were shown in Fourdrinier et al. (2011) in models where X_2 is not observed and for Kullback-Leibler loss, to be universally deficient and improved upon uniformly in terms of risk by a subclass of scale expansion variants $q_{\hat{\theta}_1,c}$ with $c-1$ positive and bounded above by a constant depending on the infimum mean squared error of $\hat{\theta}_1$. A slight adaptation of their result leads to dominating predictive densities of $\hat{q}_{\text{mle}}$, as well as other *plug–in* predictive densities which exploit the additional information $\theta_1 - \theta_2 \in A$, in terms of Kullback-Leibler risk. Similar improvements by scale expansion were obtained by Kubokawa, Marchand and Strawderman (2015, 2017) for both integrated L^1 and L^2 losses, as well as in ongoing work of LMoudden, Marchand and Kortbi for $\alpha-$divergence losses, but we will not pursue applications of these results here. This is expanded upon in Subsection 2.3.1.

(B) By duality, we mean that the frequentist risk performance of a predictive density $q_{\hat{\theta}_1,c}$ is equivalent to the point estimation frequentist risk of $\hat{\theta}_1$ in estimating θ_1 under an associated dual loss (e.g., Robert, 1996). For Kullback-Leibler risk, the dual loss is squared error (Lemma 2.3.16) and our problem connects to the problem of estimating θ_1 with $\theta_1 - \theta_2 \in A$ based on model (2.1). In turn, as expanded upon in Marchand and Strawderman (2004), improvements for the latter problem can be generated via the rotation technique (Blumenthal and Cohen, 1968, Cohen and Sackrowitz, 1970, van Eeden and Zidek, 2001, 2003) by improvements for a related restricted parameter space problem. Details are provided in Subsection 2.3.2.

Similarly, for $\alpha-$divergence loss with $\alpha \in (-1, 1)$, the predictive density risk performance of $q_{\hat{\theta}_1,c}$ connects to the point estimation frequentist risk of $\hat{\theta}_1$ in estimating θ_1, with $\theta_1 - \theta_2 \in A$ based on model (2.1), under reflected normal loss L_{γ_0} as seen in Lemma 2.3.17 below. In turn, one can capitalize on a result of Kukobawa, Marchand and Strawderman (2015) which provides a sufficient condition, expressed in terms of a dominance condition under squared error loss, for estimator $\hat{\theta}_{1,A}$ to dominate estimator $\hat{\theta}_{1,B}$ under loss L_{γ_0}. Then, proceeding as above, this latter problem connects to a restricted parameter space and analysis at this lower level provides results all the way back to the original predictive density estimation problem. Details and illustrations are provided in Subsection 2.3.2.

2.3.1 Improvements by variance expansion

Improvements on *plug–in* predictive density estimators by variance expansion stem from the following result.

Lemma 2.3.12. *Consider model (2.1) with $\theta_1 - \theta_2 \in A$, a given estimator $\hat{\theta}_1$ of θ_1, and the problem of estimating the density of Y_1 under Kullback-Leibler loss by a predictive density estimator $q_{\hat{\theta}_1,c}$ as in (2.17). Let $\underline{R} = \inf_\theta\{\mathbb{E}_\theta[\|\hat{\theta}_1(X) - \theta_1\|^2]\}/(p\sigma_Y^2)$, where the infimum is taken over the parameter space, i.e. $\{\theta \in \mathbb{R}^{2p} : \theta_1 - \theta_2 \in A\}$, and suppose that $\underline{R} > 0$.*

(a) *Then, $q_{\hat{\theta}_1,1}$ is inadmissible and dominated by $q_{\hat{\theta}_1,c}$ for $1 < c < c_0(1+\underline{R})$, with $c_0(s)$, for $s > 1$, the root $c \in (s, \infty)$ of $G_s(c) = (1 - 1/c)\, s - \log c$.*

(b) *Furthermore, we have* $s^2 < c_0(s) < e^s$ *for all* $s > 1$*, as well as* $\lim_{s\to\infty} c_0(s)/e^s = 1$.

Proof. See Fourdrinier et al. (2011, Theorem 5.1) for part **(a)**. For the first part of **(b)**, it suffices to show that **(i)** $G_s(s^2) > 0$ and **(ii)** $G_s(e^s) < 0$, given that $G_s(\cdot)$ is, for fixed s, a decreasing function on (s, ∞). We have indeed $G_s(e^s) = -se^{-s} < 0$, while $G_s(s^2)|_{s=1} = 0$ and $\frac{\partial}{\partial s} G_s(s^2) = (1 - 1/s)^2 > 0$, which implies **(i)**. Finally, set $k_0(s) = \log c_0(s), s > 1$, and observe that the definition of c_0 implies that $u(k_0(s)) = \frac{k_0(s)}{1-e^{-k_0(s)}} = s$. Since $u(k)$ increases in $k \in (1, \infty)$, it must be the case that $k_0(s)$ increases in $s \in (1, \infty)$ with $\lim_{s\to\infty} k_0(s) \geq \lim_{s\to\infty} \log s^2 = \infty$. The result thus follows since $\lim_{s\to\infty} k_0(s)/s = \lim_{s\to\infty}(1 - e^{-k_0(s)}) = 1$.

□

Remark 2.3.13. *Part (b) above is indicative of the large allowance in the degree of expansion that leads to improvement on the* plug–in *procedure. However, among these improvements* $c \in (1, c_0(1 + \underline{R}))$ *on* $q_{\hat{\theta}_1,1}$*, a complete subclass is given by the choices* $c \in [1 + \underline{R}, c_0(1 + \underline{R}))$*, while a minimal complete subclass of predictive density estimators* $q_{\hat{\theta}_1,c}$ *corresponds to the choices* $c \in [1 + \underline{R}, 1 + \overline{R}]$*, with* $\overline{R} = \sup_\theta\{\mathbb{E}_\theta[\|\hat{\theta}_1(X) - \theta_1\|^2]\}/(p\sigma_Y^2)$*, where the supremum is taken over the restricted parameter space, with* $\theta_1 - \theta_2 \in A\}$ *(see Fourdrinier et al., 2011, Remark 5.1).*

The above result is, along with Corollary 2.3.15 below, universal with respect to the choice of the *plug-in* estimator $\hat{\theta}_1$, the dimension p and the constraint set A. We will otherwise focus below on the *plug-in* maximum likelihood predictive density estimator $\hat{q}_{\text{mle}}$. The next result will be used in both this, and the following, subsections. The first part presents a decomposition of $\hat{\theta}_{1,mle}$, while the second and third parts relate to a squared error risk decomposition of estimators given by Marchand and Strawderman (2004).

Lemma 2.3.14. *Consider the problem of estimating* θ_1 *in model (2.1) with* $\theta_1 - \theta_2 \in A$ *and based on* X*. Set* $r = \sigma_2^2/\sigma_1^2$*,* $\mu_1 = (\theta_1 - \theta_2)/(1+r)$*,* $\mu_2 = (r\theta_1 + \theta_2)/(1+r)$*,* $W_1 = (X_1 - X_2)/(1 + r), W_2 = (rX_1 + X_2)/(1 + r)$*, and consider the subclass of estimators of* θ_1

$$\mathrm{C} = \{\delta_\psi : \delta_\psi(W_1, W_2) = W_2 + \psi(W_1)\} \ . \tag{2.18}$$

Then,

(a) *The maximum likelihood estimator (mle) of* θ_1 *is a member of* C *with* $\psi(W_1)$ *the mle of* μ_1 *based on* $W_1 \sim N_p(\mu_1, \sigma_1^2/(1+r)I_p)$ *and* $(1+r)\mu_1 \in A$*;*

(b) *The frequentist risk under squared error loss* $\|\delta - \theta_1\|^2$ *of an estimator* $\delta_\psi \in$ C *is equal to*

$$R(\theta, \delta_\psi) \;=\; \mathbb{E}_{\mu_1}[\|\psi(W_1) - \mu_1\|^2] + \frac{p\sigma_2^2}{1+r}\,; (1+r)\mu_1 \in A; \tag{2.19}$$

(c) *Under squared error loss, the estimator* δ_{ψ_1} *dominates* δ_{ψ_2} *iff* $\psi_1(W_1)$ *dominates* $\psi_2(W_1)$ *as an estimator of* μ_1 *under loss* $\|\psi - \mu_1\|^2$ *and the constraint* $(1+r)\mu_1 \in A$*.*

Proof. Part **(c)** follows immediately from part **(b)**. As in Marchand and Strawderman (2004), part **(b)** follows since

$$\begin{aligned} R(\theta, \delta_\psi) &= \mathbb{E}_\theta\left[\|W_2 + \psi(W_1) - \theta_1\|^2\right] \\ &= \mathbb{E}_\theta\left[\|\psi(W_1) - \mu_1\|^2\right] + \mathbb{E}_\theta\left[\|W_2 - \mu_2\|^2\right], \end{aligned}$$

yielding (2.19) given that W_1 and W_2 are independently distributed with $W_2 \sim N_p(\mu_2, (\sigma_2^2/(1+r))I_p)$. Similarly, for part **(a)**, we have $\hat{\theta}_{1,mle} = \hat{\mu}_{1,mle} + \hat{\mu}_{2,mle}$ with $\hat{\mu}_{2,mle}(W_1, W_2) = W_2$ and $\hat{\mu}_{1,mle}(W_1, W_2)$ depending only on $W_1 \sim N_p(\mu_1, (\sigma_1^2/(1+r))I_p)$ given the independence of W_1 and W_2. □

Combining Lemmas 2.3.12 and 2.3.14, we obtain the following.

Corollary 2.3.15. *Lemma 2.3.12 applies to* plug-in *predictive density estimators* $q_{\delta_\psi,1} \sim N_p(\delta_\psi, \sigma_Y^2 I_p)$ *with* $\delta_\psi \in C$*, as defined in (2.18), and*

$$\underline{R} \;=\; \frac{1}{\sigma_Y^2}\left(\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \;+\; \frac{1}{p}\,\inf_{\mu_1}\mathbb{E}[\|\psi(W_1) - \mu_1\|^2]\right). \tag{2.20}$$

Namely, $q_{\delta_\psi,c} \sim N_p(\delta_\psi, c\sigma_Y^2 I_p)$ *dominates* $q_{\delta_\psi,1}$ *for* $1 < c < c_0(1+\underline{R})$*. Moreover, we have* $c_0(1+\underline{R}) \geq (1+\underline{R})^2 \geq (1 + \frac{1}{\sigma_Y^2}\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2})^2$*. Finally, the above applies to the maximum likelihood predictive density estimator*

$$\hat{q}_{mle} \sim N_p(\hat{\theta}_{1,mle}, \sigma_Y^2 I_p)\,, \;\; \textit{with } \hat{\theta}_{1,mle}(X) = W_2 + \hat{\mu}_{1,mle}(W_1)\,, \tag{2.21}$$

and

$$\underline{R} \;=\; \frac{1}{\sigma_Y^2}\left(\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \;+\; \frac{1}{p}\,\inf_{\mu_1}\mathbb{E}[\|\hat{\mu}_{1,mle}(W_1) - \mu_1\|^2]\right), \tag{2.22}$$

where $\hat{\mu}_{1,mle}(W_1)$ *the mle of* μ_1 *based on* $W_1 \sim N_p(\mu_1, (\sigma_1^2/(1+r))I_p)$ *and under the restriction* $(1+r)\mu_1 \in A$.

With the above dominance result quite general, one further issue is the determination of the $\underline{R}$, equivalently $c_0(1+\underline{R})$, or a better lower bound. Simulation of the mean squared error in (2.20) is a possibility. Otherwise, analytically, this seems challenging, but the simple univariate order restriction case leads to the following explicit solution.

Example 2.3.1. *(Univariate case with* $\theta_1 \geq \theta_2$*)*
Consider model (2.1) with $p = 1$ *and* $A = [0, \infty)$*. The maximum likelihood predictive density estimator* $\hat{q}_{mle}$ *is given by (2.21) with* $\hat{\mu}_{1,mle}(W_1) = \max(0, W_1)$*. The mean squared error of* $\hat{\theta}_{1,mle}(X)$ *may be derived from (2.19) as equal to*

$$R(\theta, \hat{\theta}_{1,mle}) = \mathbb{E}_{\mu_1}[\,|\hat{\mu}_{1,mle}(W_1) - \mu_1|^2] + \frac{\sigma_2^2}{1+r}\,, \mu_1 \geq 0.$$

A standard calculation for the mle of a non-negative normal mean based on $W_1 \sim N\left(\mu_1, \sigma_{W_1}^2 = \sigma_1^2/(1+r)\right)$ *yields the expression*

$$\begin{aligned}\mathbb{E}_{\mu_1}[\,|\hat{\mu}_{1,mle}(W_1) - \mu_1|^2] &= \mu_1^2\,\Phi(-\frac{\mu_1}{\sigma_{W_1}}) + \int_0^\infty (w_1-\mu_1)^2\,\phi(\frac{w_1-\mu_1}{\sigma_{W_1}})\,\frac{1}{\sigma_{W_1}} dw_1 \\ &= \sigma_{W_1}^2 \left\{\frac{1}{2} + \rho^2\Phi(-\rho) + \int_0^\rho t^2\,\phi(t)\,dt\right\},\end{aligned}$$

with the change of variables $t = (w_1-\mu_1)/\sigma_{W_1}$*, and by setting* $\rho = \mu_1/\sigma_{W_1}$*. Furthermore, it is readily verified that the above risk increases in* μ_1*; as* $\frac{d}{d\rho}\left\{\rho^2\Phi(-\rho) + \int_0^\rho t^2\,\phi(t)\,dt\right\}$ $2\rho\Phi(-\rho) > 0$ *for* $\rho > 0$*, ranging from a minimum value of* $\sigma_{W_1}^2/2$ *to a supremum value of* $\sigma_{W_1}^2$*.*

Corollary 2.3.15 thus applies with

$$\underline{R} = \frac{1}{\sigma_Y^2}(\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} + \frac{\sigma_{W_1}^2}{2}) = \frac{\sigma_1^2}{\sigma_Y^2(\sigma_1^2+\sigma_2^2)}\,(\sigma_2^2+\sigma_1^2/2)\,.$$

Similarly, Remark 2.3.13 applies with $\overline{R} = \sigma_1^2/\sigma_Y^2$*.*

As a specific illustration of Corollary 2.3.15 and Remark 2.3.13, consider the equal variances case with $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2$ *for which the above yields* $\underline{R} = 3/4, \overline{R} = 1$ *and for which we can infer that:*

(a) *$q_{\hat{\theta}_{1,mle},c}$ dominates $\hat{q}_{mle}$ under Kullback-Leibler loss for $1 < c < c_0(7/4) \approx 3.48066$*

(b) *Among the class of improvements in* **(a)**, *the choices $7/4 \leq c < c_0(7/4)$ form a minimal complete subclass;*

(c) *A minimal complete subclass among the $q_{\hat{\theta}_{1,mle},c}$'s is given by the choices $c \in [1 + \underline{R}, 1 + \overline{R}] = [7/4, 2]$.*

2.3.2 Improvements through duality

We consider again here predictive density estimators $q_{\hat{\theta}_1,c}$, as in (2.17), but focus rather on the role of the plugged-in estimator $\hat{\theta}_1$. We seek improvements on benchmark choices such as $\hat{q}_{mre}$, and *plug–in* predictive densities with $c = 1$. We begin with known duality results, and namely Kullback-Leibler and reverse Kullback-Leibler losses which relate to a dual squared error loss.

Lemma 2.3.16. *For model (2.1), the frequentist risk of the predictive density estimator $q_{\hat{\theta}_1,c}$ of the density of Y_1, under both Kullback-Leibler and reverse Kullback-Leibler losses, is dual to the frequentist risk of $\hat{\theta}_1(X)$ for estimating θ_1 under squared error loss $\|\hat{\theta}_1 - \theta_1\|^2$. Namely, $q_{\hat{\theta}_{1,A},c}$ dominates $q_{\hat{\theta}_{1,B},c}$ under loss L_α iff $\hat{\theta}_{1,A}(X)$ dominates $\hat{\theta}_{1,B}(X)$ under squared error loss.*

Proof. We refer to Fourdrinier et al. (2011) for the case of Kullback-Leibler loss. For reverse Kullback-Leibler loss, the result follows as an application of Theorem 2.6.35; which is a general result for exponential families presented in the Appendix, and expanded upon with Example 2.6.5. □

For other $\alpha-$divergence losses, it is reflected normal loss (defined below) which is dual, as shown by Ghosh, Mergel and Datta (2008) for *plug–in* predictive density estimators, as well as scale expansions in (2.17).

Lemma 2.3.17. *(Duality between $\alpha-$divergence and reflected normal losses)* *For model (2.1), the frequentist risk of the predictive density estimator $q_{\hat{\theta}_1,c}$ of the density of Y_1 under $\alpha-$divergence loss (2.2), with $|\alpha| < 1$, is dual to the frequentist risk of $\hat{\theta}_1(X)$ for estimating θ_1 under reflected normal loss*

$$L_{\gamma_0}(\theta_1, \hat{\theta}_1) = 1 - e^{-\|\hat{\theta}_1 - \theta_1\|^2/2\gamma_0}\,, \tag{2.23}$$

with $\gamma_0 = (\frac{c}{1+\alpha} + \frac{1}{1-\alpha})\,\sigma_Y^2$. *Namely,* $q_{\hat{\theta}_{1,A},c}$ *dominates* $q_{\hat{\theta}_{1,B},c}$ *under loss* L_α *iff* $\hat{\theta}_{1,A}(X)$ *dominates* $\hat{\theta}_{1,B}(X)$ *under loss* L_{γ_0} *as above.*

Proof. See for instance Marchand, Perron and Yadegari (2017), or again Ghosh, Mergel and Datta (2008). □

Remark 2.3.18. *Observe that* $\lim_{\gamma_0 \to \infty} 2\gamma_0\, L_{\gamma_0}(\theta_1, \hat{\theta}_1) = \|\hat{\theta}_1 - \theta_1\|^2$, *so that the point estimation performance of* θ_1 *under reflected normal loss* L_{γ_0} *should be expected to match that of squared error loss when* $\gamma_0 \to \infty$. *In view of Lemma 2.3.16 and Lemma 2.3.17, this in turn suggests that the* $\alpha-$*divergence performance of* $\hat{q}_{\hat{\theta}_1,c}$ *will match both the Kullback-Leibler and reverse Kullback-Leibler performance when* $|\alpha| \to 1$.

Now, pairing Lemma 2.3.16 and Lemma 2.3.14 leads immediately to the following general dominance result for Kullback-Leibler and reverse Kullback-Leibler losses.

Proposition 2.3.19. *Consider model (2.1) with* $\theta_1 - \theta_2 \in A$ *and the problem of estimating the density of* Y_1 *under either Kullback-Leibler or reverse Kullback-Leibler losses. Set* $r = \sigma_2^2/\sigma_1^2$, $W_1 = (X_1 - X_2)/(1+r), W_2 = (rX_1 + X_2)/(1+r), \mu_1 = (\theta_1 - \theta_2)/(1+r)$, *and further consider the subclass of predictive densities* $q_{\delta_\psi,c}$, *as in (2.17) for fixed c, with* δ_ψ *an estimator of* θ_1 *of the form* $\delta_\psi(W_1, W_2) = W_2 + \psi(W_1)$. *Then,* $q_{\delta_{\psi_A},c}$ *dominates* $q_{\delta_{\psi_B},c}$ *if and only if* ψ_A *dominates* ψ_B *as an estimator of* μ_1 *under loss* $\|\psi - \mu_1\|^2$, *for* $W_1 \sim N_p(\mu_1, \frac{\sigma_1^2}{1+r} I_p)$ *and the parametric restriction* $(1+r)\mu_1 \in A$.

Proof. The result follows from Lemma 2.3.16 and Lemma 2.3.14. □

The above result connects three problems, namely:

- **(I)** the efficiency of $q_{\delta_\psi,c}$ under KL or RKL loss as a predictive density for Y_1 with the additional information $\theta_1 - \theta_2 \in A$;
- **(II)** the efficiency of $\delta_\psi(X)$ as an estimator of θ_1 under squared error loss $\|\delta_\psi - \theta_1\|^2$ with the additional information $\theta_1 - \theta_2 \in A$;
- **(III)** the efficiency of $\psi(W_1)$ for $W_1 \sim N_p(\mu_1, \sigma_1^2/(1+r)I_p)$ as an estimator of μ_1 under squared error loss $\|\psi - \mu_1\|^2$ with the parametric restriction $(1+r)\mu_1 \in A$.

Previous authors (Blumenthal and Cohen, 1968; Cohen and Sackrowitz, 1970; van Eeden and Zidek, (2001, 2003), for $p = 1$; Marchand and Strawderman, 2004, for $p \geq 1$) have exploited the **(II)-(III)** connection (i.e., Lemma 2.3.14) to obtain findings for problem **(II)** based on restricted parameter space findings for **(III)**. The above Proposition further exploits connections **(I)-(II)** (i.e., Lemma 2.3.16) to derive findings for predictive density estimation problem **(I)** from restricted parameter space findings for **(III)**. Consequently, findings for **(III)-(II)** provide findings for our predictive density estimation problem **(I)**, and we refer for Marchand and Strawderman (2004), as well as the references therein, for examples of such results. An example, which is also illustrative of $\alpha-$divergence results, is provided below at the end of this section.

For $\alpha-$divergence losses other than Kullback-Leibler and reverse Kullback-Leibler, the above scheme is not immediately available for the dual reflected normal loss since Lemma 2.3.14 is intimately linked to squared error loss. However, a slight extension of Lemma 3.3 of Kubokawa, Marchand and Strawderman (2015); exploiting a concave loss technique dating back to Brandwein and Strawderman (1980); permits us to connect (but only in one direction) reflected normal loss to squared error loss, and consequently the efficiency of predictive densities under α-divergence loss to point estimation in restricted parameter spaces as in **(III)** above.

Lemma 2.3.20. *Consider model (2.1) and the problem of estimating θ_1 based on X, with $\theta_1 - \theta_2 \in A$ and reflected normal loss as in (2.23) with $|\alpha| < 1$. Then $\hat{\theta}_1(X)$ dominates X_1 whenever $\hat{\theta}_1(Z)$ dominates Z_1 as an estimate of θ_1, under squared error loss $\|\hat{\theta}_1 - \theta_1\|^2$, with $\theta_1 - \theta_2 \in A$, for the model*

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathrm{N}_{2p}\left(\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \Sigma_Z = \begin{pmatrix} \sigma^2_{Z_1} I_p & 0 \\ 0 & \sigma_2^2 I_p \end{pmatrix}\right), \tag{2.24}$$

with $\sigma^2_{Z_1} = \frac{\gamma \sigma_1^2}{\gamma + \sigma_1^2}$.

Proof. Denote the loss $\rho(\|\hat{\theta}_1 - \theta_1\|^2)$ with $\rho(t) = 1 - e^{-t/2\gamma}$. Since ρ is concave, we have for all $x = (x_1, x_2)' \in \mathbb{R}^{2p}$:

$$\rho(\|\hat{\theta}_1(x) - \theta_1\|^2) - \rho(\|x_1 - \theta_1\|^2) \leq \rho'(\|x_1 - \theta_1\|^2)\left(\|\hat{\theta}_1(x) - \theta_1\|^2 - \|x_1 - \theta_1\|^2\right).$$

With $\rho'(t) = \frac{1}{2\gamma} e^{-t/2\gamma}$, we have for the difference in risks and $Z \sim f_Z$:

$$\begin{aligned}\Delta(\theta) &= R(\theta, \hat{\theta}_1) - R(\theta, X_1) \\ &\leq \frac{1}{2\gamma} \frac{1}{(2\pi\sigma_1\sigma_2)^p} \int_{\mathbb{R}^{2p}} e^{-\frac{\|x_1-\theta_1\|^2}{2\gamma}} \left(\|\hat{\theta}_1(x) - \theta_1\|^2 - \|x_1 - \theta_1\|^2 \right) e^{-\frac{\|x_1-\theta_1\|^2}{2\sigma_1^2} - \frac{\|x_2-\theta_2\|^2}{2\sigma_2^2}} dx \\ &= \frac{1}{2\gamma} (\frac{\gamma}{\gamma + \sigma_1^2})^{p/2} \int_{\mathbb{R}^{2p}} \left(\|\hat{\theta}_1(z) - \theta_1\|^2 - \|z_1 - \theta_1\|^2 \right) f_Z(z) dz \,,\end{aligned}$$

establishing the result. □

Proposition 2.3.21. *Consider model (2.1) with $\theta_1 - \theta_2 \in A$ and the problem of estimating the density of Y_1 under either Kullback-Leibler or reverse Kullback-Leibler losses. Set $r = \sigma_2^2/\sigma_1^2$, $W_1 = (X_1 - X_2)/(1+r), W_2 = (rX_1 + X_2)/(1+r), \mu_1 = (\theta_1 - \theta_2)/(1+r)$, and further consider the subclass of predictive densities $q_{\delta_\psi, c}$, as in (2.17) for fixed c, with δ_ψ an estimator of θ_1 of the form $\delta_\psi(W_1, W_2) = W_2 + \psi(W_1)$. Then, $q_{\delta_{\psi_A}, c}$ dominates $q_{\delta_{\psi_B}, c}$ as long as ψ_A dominates ψ_B as an estimator of μ_1 under loss $\|\psi - \mu_1\|^2$, for $W_1 \sim N_p(\mu_1, \frac{\sigma_{Z_1}^2}{1+r} I_p)$, the parametric restriction $(1+r)\mu_1 \in A$, and $\sigma_{Z_1}^2 = \frac{\{(1+\alpha)+c(1-\alpha)\}\,\sigma_1^2}{\{(1+\alpha)+c(1-\alpha)\} + (1-\alpha^2)\,\sigma_1^2/\sigma_Y^2}$.*

Proof. The result follows from Lemma 2.3.17 and its dual reflected normal loss L_{γ_0}, the use of Lemma 2.3.20 applied to $\sigma_{Z_1}^2 = \frac{\gamma_0 \sigma_1^2}{\gamma_0 + \sigma_1^2}$, and an application of part **(c)** of Lemma 2.3.14 to Z as distributed in (2.24). □

Remark 2.3.22. *Proposition 2.3.21 holds as stated for $|\alpha| = 1$ and is thus a continuation of the sufficiency part of Proposition 2.3.19. As well, the above result provides positive findings as long as ψ_B is inadmissible under squared error loss and dominating estimators ψ_A are available. Many particular cases follow from the above. These include: (i) Hellinger loss with $\alpha = 0$ and $\sigma_{Z_1}^2$ simplifying to $\{(c+1)/(c+1+\sigma_1^2/\sigma_Y^2)\}\ \sigma_1^2$; (ii)* plug-in *predictive densities with $c = 1$; (iii) cases where $q_{\delta_{\psi_B}} \equiv \hat{q}_{mre}$ with the corresponding choice $c = 1 + \frac{(1-\alpha)\sigma_1^2}{2\sigma_Y^2}$ yielding*

$$\sigma_{Z_1}^2 = \frac{4\sigma_Y^2 + (1-\alpha)^2\sigma_1^2}{4\sigma_Y^2 + (3+\alpha)(1-\alpha)\,\sigma_1^2}\ \sigma_1^2 \,.$$

The above $\alpha-$ divergence result connects four problems, namely:

- **(I)** the efficiency of $q_{\delta_\psi, c}$ under $\alpha-$divergence loss, $-1 < \alpha < 1$, as a predictive density for Y_1 with the additional information $\theta_1 - \theta_2 \in A$;

(IB) the efficiency of $\delta_\psi(X)$ as an estimator of θ_1 under reflected normal loss L_{γ_0} with $\gamma_0 = (\frac{c}{1+\alpha} + \frac{1}{1-\alpha})\,\sigma_Y^2$ with the additional information $\theta_1 - \theta_2 \in A$;

(II) the efficiency of $\delta_\psi(Z)$, for Z distributed as in (2.24) with $\sigma_{Z_1}^2 = (\gamma_0\sigma_1^2)/(\gamma_0 + \sigma_1^2)$, as an estimator of θ_1 under squared error loss $\|\delta_\psi - \theta_1\|^2$ with the additional information $\theta_1 - \theta_2 \in A$;

(III) the efficiency of $\psi(W_1)$ for $W_1 \sim N_p(\mu_1, \sigma_{Z_1}^2/(1+r)I_p)$ as an estimator of μ_1 under squared error loss $\|\psi - \mu_1\|^2$ with the parametric restriction $(1+r)\mu_1 \in A$.

Example 2.3.2. *Here is an illustration of both Propositions 2.3.19 and 2.3.21. Consider model (2.1) with A a convex set with a non-empty interior, and $\alpha-$divergence loss ($|\alpha| \leq 1$) for assessing a predictive density for Y_1. Further consider the minimum risk predictive density $\hat{q}_{mre}$ as a benchmark procedure, which is of the form $q_{\delta_{\psi_B}}$ as in Proposition 2.3.21 with $\delta_{\psi_B} \in C$, $\psi_B(W_1) = W_1$ and $c = c_{mre} = 1 + (1-\alpha)\sigma_1^2/(2\sigma_Y^2)$. Now consider the Bayes estimator $\psi_U(W_1)$ under squared error loss of μ_1 associated with a uniform prior on the restricted parameter space $(1+r)\mu_1 \in A$, for $W_1 \sim N_p((\mu_1, \frac{\sigma_{Z_1}^2}{1+r} I_p)$ as in Proposition 2.3.21. It follows from Hartigan's theorem (Hartigan, 2003; Marchand and Strawderman, 2004) that $\psi_A(W_1) \equiv \psi_U(W_1)$ dominates $\psi_B(W_1)$ under loss $\|\psi - \mu_1\|^2$ and for $(1+r)\mu_1 \in A$. It thus follows from Proposition 2.3.21 that the predictive density $q_{\delta_{\psi_B}, c_{mre}} \sim N_p(\delta_{\psi_B}(X), (\frac{1-\alpha}{2}\sigma_1^2 + \sigma_Y^2)I_p)$ dominates $\hat{q}_{mre}$ under $\alpha-$divergence loss with $\delta_{\psi_B}(X) = \frac{rX_1+X_2}{1+r} + \psi_U(\frac{X_1-X_2}{1+r})$. The dominance result is unified with respect to $\alpha \in [-1,1]$, the dimension p, and the set A.*

We conclude this section with an adaptive two-step strategy, building on both variance expansion and improvements through duality, to optimise on potential Kullback-Leibler improvements on a maximum likelihood estimator predictive density estimator in model (2.1) of the form $\hat{q}_{mle} \sim N_p(\hat{\theta}_{1,mle}, \sigma_Y^2 I_p)$, in cases where point estimation improvements on $\hat{\theta}_{1,mle}(X)$ under squared error loss are readily available.

(I) Select an estimator $\hat{\theta}_1^*$ which dominates $\hat{\theta}_{1,mle}$ under squared error loss. This may be achieved via part **(c)** of Lemma 2.3.14 resulting in a dominating estimator of the form $\hat{\theta}_1^*(X) = W_2 + \psi^*(W_1) = (rX_1 + X_2)/(1+r) + \psi^*((X_1 - X_2)/(1+r))$ where $\psi^*(W_1)$ dominates $\hat{\mu}_{1,mle}(W_1)$ as an estimator of μ_1 under squared error loss and the restriction $(1+r)\mu_1 \in A$.

(II) Now, with the *plug-in* predictive density estimator $q_{\hat{\theta}_{1*},1}$ dominating $\hat{q}_{mle}$, further improve $q_{\hat{\theta}_{1*},1}$ by a variance expanded $q_{\hat{\theta}_{1*},c}$. Suitable choices of c are prescribed by Corollary 2.3.15 and given by $c_0(1+\underline{R})$, with $\underline{R}$ given in (2.20). The evaluation of $\underline{R}$ hinges on the infimum risk $\inf_{\mu_1} \mathbb{E}[\|\psi^*(W_1) - \mu_1\|^2]$, and such a quantity can be either estimated by simulation, derived in some cases analytically, or safely underestimated by 0.

Examples where the above can be applied include the cases: **(i)** $A = [0, \infty)$ with the use of Shao and Strawderman's (1996) dominating estimators, and **(ii)** A the ball of radius m centered at the origin with the use of Marchand and Perron's (2001) dominating estimators. [2]

2.4 Bayesian dominance results

In the previous section, we studied the efficiency of predictive densities as in (2.17) and elaborated on methods to obtain improvements, whenever possible, for instance on *plug-in* and minimum risk equivariant predictive density estimators. We focus here on Bayesian improvements, for reverse Kullback-Leibler and Kullback-Leibler losses, of the benchmark minimum risk equivariant predictive density estimator. For Kullback-Leibler loss, we establish that the uniform Bayes predictive density estimator $\hat{q}_{\pi_{U,A}}$ dominates $\hat{q}_{mre}$ for the univariate cases where $\theta_1 - \theta_2$ is either restricted to a compact interval, lower-bounded or upper-bounded. Our findings for reverse Kullback-Leibler loss are more wide ranging. Indeed, we exploit the fact that Bayes predictive density estimators are *plug–in* predictive density estimators, that the comparison of such procedures is dual to point estimation comparisons under squared error loss, and that we thus can capitalize on existing results for our purposes via Lemma 2.3.14. Such properties are, as expanded upon in the Appendix, quite general for exponential families and reverse Kullback-Leibler loss.

[2] Alternatively, one could expand the variance first, and then improve on the *plug-in*; such as using a Shao and Strawderman estimator to obtain an improvement on $q_{\hat{\theta}_{1,mle},c}$ in Example 2.3.1; but this may be suboptimal in view of the complete class considerations of Remark 2.3.13.

2.4.1 Reverse Kullback-Leibler loss

We begin with an identification of Bayes predictive densities that belong to the class C given in (2.18), which will permit us to apply Lemma 2.3.14 in decomposing the frequentist risk of such procedures. This formalizes and extends representation (2.16).

Lemma 2.4.23. *Consider model (2.1) and the problem of estimating θ_1 based on X with $\theta_1 - \theta_2 \in A$ and loss $\|\delta - \theta_1\|^2$. Set $r = \sigma_2^2/\sigma_1^2$, $\mu_1 = (\theta_1 - \theta_2)/(1+r), \mu_2 = (r\theta_1 + \theta_2)/(1+r)$, $W_1 = (X_1 - X_2)/(1+r), W_2 = (rX_1 + X_2)/(1+r)$, and consider prior densities of the form $\pi(\theta) = \pi_1(\mu_1)\,\mathbb{I}_A((1+r)\mu_1)\,\mathbb{I}_{\mathbb{R}^p}(\mu_2)$. Then, the corresponding Bayes estimators $\hat{\theta}_{1,\pi}$ are members of the subclass C, as defined in (2.18), and are given by*

$$\hat{\theta}_{1,\pi}(X) = \psi_\pi(W_1) + W_2\,, \tag{2.25}$$

where $\psi_\pi(W_1)$ is the Bayes estimator based on $W_1 \sim N_p(\mu_1, \frac{\sigma_1^2}{1+r} I_p)$ of μ_1 for loss $\|\psi - \mu_1\|^2$ and prior $\pi_1(\mu_1)\,\mathbb{I}_A((1+r)\mu_1)$.

Proof. The result follows since the Bayes point estimator of θ_1 is given by $\mathbb{E}(\theta_1|x) = \mathbb{E}(\mu_1|w_1,w_2) + \mathbb{E}(\mu_2|w_1,w_2) = \mathbb{E}(\mu_1|w_1) + \mathbb{E}(\mu_2|w_2) = \psi_\pi(w_1) + w_2$, given the independence of W_1, W_2 and the multiplicative aspect of the prior which imply $\mu_1|w_1,w_2 =^d \mu_1|w_1$ and $\mu_2|w_1,w_2 =^d \mu_2|w_1$. □

Proposition 2.4.24. *Consider model (2.1) with $\theta_1 - \theta_2 \in A$, a prior density of the form $\pi(\theta) = \pi_1(\mu_1)\,\mathbb{I}_A((1+r)\mu_1)$, and the corresponding Bayes predictive density $\hat{q}_\pi$ for estimating the density of Y_1 under reverse Kullback-Leibler loss. Set $r = \sigma_2^2/\sigma_1^2$, $W_1 = (X_1 - X_2)/(1+r), W_2 = (rX_1 + X_2)/(1+r), \mu_1 = (\theta_1 - \theta_2)/(1+r)$, and let $q_{\delta_{\psi_0}}(\cdot;X) \sim N_p(\delta_{\psi_0}(X), \sigma_Y^2 I_p)$ be a competing* plug-in *predictive density with $\delta_{\psi_0} \in C$ of the form $\delta_\psi(W_1,W_2) = \psi_0(W_1) + W_2$. Then, $\hat{q}_\pi(\cdot;X)$ dominates $q_{\delta_{\psi_0}}(\cdot;X)$ if and only if the Bayes estimator $\psi_\pi(W_1)$, with respect to the prior $\pi_1(\mu_1)\,\mathbb{I}_A((1+r)\mu_1)$, dominates $\psi_0(W_1)$ as an estimator of μ_1 under loss $\|\psi - \mu_1\|^2$, for $W_1 \sim N_p(\mu_1, \frac{\sigma_1^2}{1+r} I_p)$ and $(1+r)\mu_1 \in A$.*

Proof. Part **(b)** of Lemma 2.2.3 and Lemma 2.4.23 tell us that $\hat{q}_\pi$ is a plug-in predictive density of the form $N_p(\hat{\theta}_{1,\pi}(X), \sigma_Y^2 I_p)$ with $\hat{\theta}_{1,\pi}(X)$ as in (2.25). In turn, Lemma 2.3.16 implies that the reverse Kullback-Leibler risk comparison between

$\hat{q}_\pi$ and $q_{\delta_{\psi_0}}$ hinges on the mean squared error comparison between $\hat{\theta}_{1,\pi}$ and δ_{ψ_0} under model (2.1). Finally, the result follows by making use of Lemma 2.3.14. □

We pursue with applications.

Example 2.4.3. *Consider the context of Proposition 2.4.24 with A a convex set with a non-empty interior, the restricted to A uniform prior $\pi_{U,A}(\theta) = \mathbb{I}_A(\theta_1 - \theta_2)$ and its corresponding Bayes predictive density $\hat{q}_{\pi_{U,A}}$ (see section 2.2.4.), and the minimum risk predictive density $\hat{q}_{mre}(\cdot;X) \sim N_p(X_1, \sigma_Y^2 I_p)$. It follows from Hartigan's theorem that the Bayes estimator $\psi_U(W_1)$ dominates $\psi_0(W_1) = W_1$ under squared error loss. Hence, from Proposition 2.4.24, it follows that the Bayes predictive density $\hat{q}_{\pi_{U,A}}$ dominates $\hat{q}_{mre}$ for reverse Kullback-Leibler loss. The result is general with respect to the choices of p and A.*

For $p = 1$ and $A = [-m, m]$, Kubokawa (2005), as well as Marchand and Payandeh (2011), provide alternative Bayes estimators $\psi_{\pi_a}(W_1)$ which dominate as well W_1 for priors π_a supported on the set $\mu_1 \in [\frac{-m}{1+r}, \frac{m}{1+r}]$. In turn, and as above for the uniform prior, it thus follows that the corresponding Bayes predictive densities $\hat{q}_\pi(\cdot;X) \sim N(\psi_{\pi_a}(W_1) + W_2, \sigma_Y^2)$ dominate $\hat{q}_{mre}$ with $\pi(\theta) = \pi_a(\mu_1)\mathbb{I}_{\mathbb{R}}(\mu_2)$.

Remark 2.4.25. *For $p \geq 3$, $\hat{q}_{\pi_U}$, as well as* plug-in *predictive density of the form $q_{\delta_{\psi_0}}(\cdot;X) \sim N_p(\psi_0(W_1) + W_2, \sigma_Y^2 I_p)$, are inadmissible and dominated by predictive densities $q_{\delta_{\psi_0,\psi_1}}(\cdot;X) \sim N_p(\psi_0(W_1) + \psi_1(W_2), \sigma_Y^2 I_p)$ where $\psi_1(W_2)$ is an estimator of μ_2, for $W_2 \sim N_p(\mu_2, \frac{\sigma_2^2}{1+r} I_p)$, which dominates W_2. Stein estimation findings (e.g., Stein, 1981) provide many such dominating estimators, including Bayesian improvements. For instance, for $p \geq 3$ and a superharmonic prior π_2 for μ_2, the predictive density $\hat{q}_{\pi_U}$ is dominated by the Bayes predictive density $q_{\delta_{\psi_U,\psi_{\pi_2}}}(\cdot;X) \sim N_p(\psi_U(W_1) + \psi_{\pi_2}(W_2), \sigma_Y^2 I_p)$, associated with the prior $\pi(\theta) = \mathbb{I}_A((1+r)\mu_1)\, \pi_2(\mu_2)$. The above inferences come about a rewriting of Lemma 2.3.14 for estimators of the form $\psi_0(W_1) + \psi_1(W_2)$, with $\psi_0 \equiv \psi_U$ for the case of $\hat{q}_{mre}$ and its use as in Proposition 2.4.24.*

Example 2.4.4. *Consider the context of Proposition 2.4.24 and the maximum likelihood predictive density estimator $\hat{q}_{mle} \sim N_p(\hat{\theta}_{1,mle}, \sigma_Y^2 I_p)$ with $\hat{\theta}_{1,mle}(X) = W_2 + \psi_0(W_1)$, as in (2.21) with $\psi_0(W_1) = \hat{\mu}_{1,mle}(W_1)$. It follows from Lemma 2.3.16 that* plug-in *predictive densities $N_p(\psi_1(W_1) + W_2, \sigma_Y^2 I_p)$ dominate $\hat{q}_{mle}$ under reverse Kullback-Leibler loss if and only if $\psi_1(W_1)$ dominates $\hat{\mu}_{1,mle}(W_1)$ under squared error loss. In particular and in accordance with Proposition 2.4.24,* *a Bayes predictive density $\hat{q}_\pi$, for prior $\pi(\theta) = \pi_1(\mu_1)\,\mathbb{I}_A((1+r)\mu_1)\mathbb{I}_{\mathbb{R}^p}(\mu_2)$, dominates $\hat{q}_{mle}$ if and only if $\psi_\pi(W_1)$ dominates $\hat{\mu}_{1,mle}(W_1)$, where $\psi_\pi(W_1)$ is the Bayes point estimator of μ_1 for prior $\pi_1(\mu_1)\,\mathbb{I}_A((1+r)\mu_1)$. The determination of such dominating Bayesian ψ_π is challenging though. For the specific case of A being a p-dimensional ball of radius m centered at the origin, Marchand and Perron (2001), as well as Fourdrinier and Marchand (2010), provide several applicable Bayesian dominance results.*

2.4.2 Kullback-Leibler loss

In this subsection, we show, for $\theta_1 - \theta_2$ either lower bounded, upper bounded, or bounded to an interval, that the uniform Bayes predictive density estimator $\hat{q}_{\pi_{U,A}}$ improves uniformly on the minimum risk equivariant predictive density estimator $\hat{q}_{\text{mre}}$ under Kullback-Leibler loss. Without loss of generality (given Remark 2.1.1), we consider the restrictions $\theta_1 \geq \theta_2$ and $|\theta_1 - \theta_2| \leq m$. We also investigate situations where the variances of model (2.1) are misspecified, but where the dominance persists. We begin with the lower bounded case.

Theorem 2.4.26. *Consider model (2.1) with $p = 1$ and $A = [0, \infty)$. For estimating the density of Y_1 under Kullback-Leibler loss, the Bayes predictive density $\hat{q}_{\pi_{U,A}}$ dominates the minimum risk equivariant predictive density estimator $\hat{q}_{\text{mre}}$. The Kullback-Leibler risks are equal iff $\theta_1 = \theta_2$.*

Proof. Making use of Corollary 2.2.10's representation of $\hat{q}_{\pi_{U,A}}$, the difference in risks is given by

$$\begin{aligned}
\Delta(\theta) &= R_{KL}(\theta, \hat{q}_{\text{mre}}) - R_{KL}(\theta, \hat{q}_{\pi_{U,A}}) \\
&= \mathbb{E}^{X,Y_1} \log\left(\frac{\hat{q}_{\pi_{U,A}}(Y_1; X)}{\hat{q}_{\text{mre}}(Y_1; X)}\right) \\
&= \mathbb{E}^{X,Y_1} \log\left(\Phi(\alpha_0 + \alpha_1 \frac{Y_1 - X_1}{\tau})\right) - \mathbb{E}^{X,Y_1} \log\left(\Phi(\frac{\alpha_0}{\sqrt{1+\alpha^2}}\right), \qquad (2.26)
\end{aligned}$$

with $\alpha_0 = \frac{X_1 - X_2}{\sigma_T}$, $\alpha_1 = \frac{\beta\tau}{\sigma_T}$, $\tau = \sqrt{\sigma_1^2 + \sigma_Y^2}$, $\beta = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_Y^2}$, and $\sigma_T^2 = \sigma_2^2 + \beta\sigma_Y^2$. Now, observe that

$$\alpha_0 + \alpha_1 \frac{Y_1 - X_1}{\tau} = \frac{1}{\sigma_T}(X_1 - X_2 + \beta(Y_1 - X_1)) \sim N(\frac{\theta_1 - \theta_2}{\sigma_T}, 1)\,, \qquad (2.27)$$

and

$$\frac{\alpha_0}{\sqrt{1+\alpha_1^2}} = \frac{X_1 - X_2}{\sqrt{\sigma_1^2+\sigma_2^2}} \sim N(\frac{\theta_1-\theta_2}{\sqrt{\sigma_1^2+\sigma_2^2}}, 1)\,. \tag{2.28}$$

We thus can write

$$\Delta(\theta) \;=\; \mathbb{E}\, G(Z)\,,$$

$$\text{with } G(Z) = \log\Phi(Z + \frac{\theta_1-\theta_2}{\sigma_T}) - \log\Phi(Z + \frac{\theta_1-\theta_2}{\sqrt{\sigma_1^2+\sigma_2^2}})\,, Z \sim N(0,1)\,.$$

With $\theta_1 - \theta_2 \geq 0$ and $\sigma_T^2 < \sigma_1^2 + \sigma_2^2$, we infer that $\mathbb{P}_\theta(G(Z) \geq 0) = 1$ and $\Delta(\theta) \geq 0$ for all θ such that $|\theta_1 - \theta_2| \leq m$, with equality iff $\theta_1 - \theta_2 = 0$. □

We now obtain an analogue dominance result in the univariate case for the additional information $\theta_1 - \theta_2 \in [-m, m]$.

Theorem 2.4.27. *Consider model (2.1) with $p = 1$ and $A = [-m, m]$. For estimating the density of Y_1 under Kullback-Leibler loss, the Bayes predictive density $\hat{q}_{\pi_{U,A}}$ (strictly) dominates the minimum risk equivariant predictive density estimator $\hat{q}_{mre}$.*

Proof. Making use of (2.12) and (2.13) for the representation of $\hat{q}_{\pi_{U,A}}$, the difference in risks is given by

$$\begin{aligned}
\Delta(\theta) &= R_{KL}(\theta, \hat{q}_{\text{mre}}) - R_{KL}(\theta, \hat{q}_{\pi_{U,A}}) \\
&= \mathbb{E}^{X,Y_1} \log\left(\frac{\hat{q}_{\pi_{U,A}}(Y_1;X)}{\hat{q}_{\text{mre}}(Y_1;X)}\right) \\
&= \mathbb{E}^{X,Y_1} \log\left(\Phi(\alpha_0 + \alpha_1\frac{Y_1-X_1}{\tau}) - \Phi(\alpha_2 + \alpha_1\frac{Y_1-X_1}{\tau})\right) \\
&- \mathbb{E}^{X,Y_1} \log\left(\Phi(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}}) - \Phi(\frac{\alpha_2}{\sqrt{1+\alpha_1^2}})\right)\,,
\end{aligned}$$

with the α_i's given in Section 2.2. Now, observe that

$$\alpha_0 + \alpha_1\frac{Y_1-X_1}{\tau} = \frac{1}{\sigma_T}(m + (X_1-X_2) + \beta(Y_1-X_1) \sim N(\delta_0 = \frac{m+\theta_1-\theta_2}{\sigma_T}, 1)\,, \tag{2.29}$$

and

$$\frac{\alpha_0}{\sqrt{1+\alpha_1^2}} = \frac{(m+(X_1-X_2)}{\sqrt{\sigma_1^2+\sigma_2^2}} \sim N(\delta_0' = \frac{m+\theta_1-\theta_2}{\sqrt{\sigma_1^2+\sigma_2^2}}, 1)\,. \tag{2.30}$$

Similarly, we have $\alpha_2 + \alpha_1 \frac{Y_1 - X_1}{\tau} \sim N(\delta_2 = \frac{-m+\theta_1-\theta_2}{\sigma_T}, 1)$ and $\frac{\alpha_2}{\sqrt{1+\alpha_1^2}} \sim N(\delta_2' = \frac{-m+\theta_1-\theta_2}{\sqrt{\sigma_1^2+\sigma_2^2}}, 1)$. We thus can write

$$\Delta(\theta) \,=\, \mathbb{E}H(Z)\,,$$

with $H(Z) = \log\left(\Phi(Z+\delta_0) - \Phi(Z+\delta_2)\right) - \log\left(\Phi(Z+\delta_0') - \Phi(Z+\delta_2')\right)\,, Z \sim N(0,1)\,.$

With $-m \leq \theta_1 - \theta_2 \leq m$ and $\sigma_T^2 < \sigma_1^2 + \sigma_2^2$, we infer that $\delta_0 \geq \delta_0'$ with equality iff $\theta_1 - \theta_2 = -m$ and $\delta_2 \leq \delta_2'$ with equality iff $\theta_1 - \theta_2 = m$, so that $\mathbb{P}_\theta(H(Z) > 0) = 1$ and $\Delta(\theta) > 0$ for all θ such that $|\theta_1 - \theta_2| \leq m$. □

We now investigate situations where the variances in model (2.1) are misspecified. To this end, we consider σ_1^2, σ_2^2 and σ_Y^2 as the nominal variances used to construct the predictive density estimates $\hat{q}_{\pi_{U,A}}$ and $\hat{q}_{\text{mre}}$, while the true variances, used to assess frequentist Kullback-Leibler risk, are, unbeknownst to the investigator, given by $a_1^2\sigma_1^2$, $a_2^2\sigma_2^2$ and $a_Y^2\sigma_Y^2$ respectively. We exhibit, below in Theorem 2.4.30, many combinations of the nominal and true variances such that the Theorem 2.4.26's dominance result persists. Such conditions for the dominance to persist includes the case of equal a_1^2, a_2^2 and a_Y^2 (i.e., the three ratios true variance over nominal variance are the same), among others.

We require the following intermediate result.

Lemma 2.4.28. *Let $U \sim N(\mu_U, \sigma_U^2)$ and $V \sim N(\mu_V, \sigma_V^2)$ with $\mu_U \geq \mu_V$ and $\sigma_U^2 \leq \sigma_V^2$. Let H be a differentiable function such that both H and $-H'$ are increasing. Then, we have $\mathbb{E}H(U) \geq \mathbb{E}H(V)$.*

Proof. Suppose without loss of generality that $\mu_V = 0$, and set $s = \frac{\sigma_U}{\sigma_V}$. Since U and $\mu_U + sV$ share the same distribution and $\mu_U \geq 0$, we have:

$$\begin{aligned}
\mathbb{E}H(U) &= \mathbb{E}H(\mu_U + sV) \\
&\geq \mathbb{E}H(sV) \\
&= \int_{\mathbb{R}_+} \left(H(sv) + H(-sv)\right) \frac{1}{\sigma_V}\phi(\frac{v}{\sigma_V})\, dv\,.
\end{aligned}$$

Differentiating with respect to s, we obtain

$$\frac{d}{ds}\,\mathbb{E}H(sV) \,=\, \int_{\mathbb{R}_+} v\,\left(H'(sv) - H'(-sv)\right) \frac{1}{\sigma_V}\phi(\frac{v}{\sigma_V})\, dv \,\leq 0$$

since H' is decreasing. We thus conclude that

$$\mathbb{E}H(U) \geq \mathbb{E}H(sV) \geq \mathbb{E}H(V)\,,$$

since $s \leq 1$ and H is increasing by assumption. □

Remark 2.4.29. *We point out that the result and proof extend to location-scale families* $U \sim \frac{1}{\sigma_U} f_0(\frac{t-\mu_U}{\sigma_U})$,
$V \sim \frac{1}{\sigma_V} f_0(\frac{t-\mu_V}{\sigma_V})$, *with even* f_0, $\mu_U \geq \mu_V$, *and* $\sigma_U < \sigma_V$.

Theorem 2.4.30. *Consider model (2.1) with* $p = 1$ *and* $A = [0, \infty)$. *Suppose that the variances are misspecified and that the true variances are given by* $\mathbb{V}(X_1) = a_1^2\sigma_1^2, \mathbb{V}(X_2) = a_2^2\sigma_2^2, \mathbb{V}(Y_1) = a_Y^2\sigma_Y^2$. *For estimating the density of* Y_1 *under Kullback-Leibler loss, the Bayes predictive density* $\hat{q}_{\pi_{U,A}}$ *dominates the minimum risk equivariant predictive density estimator* $\hat{q}_{mre}$ *whenever* $\sigma_U^2 \leq \sigma_V^2$ *with*

$$\sigma_U^2 = \frac{a_2^2\sigma_2^2 + (1-\beta)^2 a_1^2\sigma_1^2 + \beta^2 a_Y^2\sigma_Y^2}{\sigma_2^2 + \beta\sigma_Y^2}\,,\, \sigma_V^2 = \frac{a_1^2\sigma_1^2 + a_2^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\,, \beta = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_Y^2}. \tag{2.31}$$

In particular, dominance occurs for cases : **(i)** $a_1^2 = a_2^2 = a_Y^2$, **(ii)** $a_Y^2 \leq a_1^2 = a_2^2$, **(iii)** $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2$ *and* $\frac{a_2^2 + a_Y^2}{2} \leq a_1^2$.

Remark 2.4.31. *Conditions* **(i)**, **(ii)** *and* **(iii)** *are quite informative. One common factor for the dominance to persist, especially seen by* **(iii)**, *is for the variance of* X_1 *to be relatively large compared to the variances of* X_2 *and* Y_1.

Proof. Particular cases **(i)**, **(ii)**, **(iii)** follow easily from (2.31). To establish condition (2.31), we prove, as in Theorem 2.4.26, that $\Delta(\theta)$ given in (2.26) is greater or equal to zero. We apply Lemma 2.4.28, with $H \equiv \log \Phi$ increasing and concave as required, showing that $\mathbb{E}\, log\Phi(U) \geq \mathbb{E}\, log\Phi(V)$ with $U = \alpha_0 + \alpha_1 \frac{Y_1 - X_1}{\tau} \sim N(\mu_U, \sigma_U^2)$ and $V = \frac{\alpha_0}{\sqrt{1+\alpha_1^2}} \sim N(\mu_V, \sigma_V^2)$. Since $\mu_U = \frac{\theta_1 - \theta_2}{\sigma_T} > \frac{\theta_1 - \theta_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \mu_V$, the inequality $\sigma_U^2 \leq \sigma_V^2$ will suffice to have dominance. Finally, the proof is complete by checking that σ_U^2 and σ_V^2 are as given in (2.31), when the true variances are given by $\mathbb{V}(X_1) = a_1^2\sigma_1^2, \mathbb{V}(X_2) = a_2^2\sigma_2^2, \mathbb{V}(Y_1) = a_Y^2\sigma_Y^2$. □

Remark 2.4.32. *In opposition to the above robustness analysis, the dominance property of* $\hat{q}_{\pi_{U,A}}$ *versus* $\hat{q}_{mre}$ *for the restriction* $\theta_1 - \theta_2 \geq 0$ *does not persists for parameter space values such that* $\theta_1 - \theta_2 < 0$, *i.e., the additional information difference is misspecified. In fact, it is easy to see following the proof of Theorem 2.4.26 that* $R_{KL}(\theta, \hat{q}_{mre}) - R_{KL}(\theta, \hat{q}_{\pi_{U,A}}) < 0$ *for* θ*'s such that* $\theta_1 - \theta_2 < 0$. *A* *potential protection is to use the predictive density estimator* $\hat{q}_{\pi_{U,A'}}$ *with* $A' = [\epsilon, \infty)$, $\epsilon < 0$, *and with dominance occurring for all* θ *such that* $\theta_1 - \theta_2 \geq \epsilon$ *(Remark 2.1.1 and Theorem 2.4.26).*

2.5 Examples, illustrations and further comments

We present and comment numerical evaluations of Kullback-Leibler risks in the univariate case for both $\theta_1 \geq \theta_2$ (Figures 2.1, 2.2) and $|\theta_1 - \theta_2| \leq m, m = 1, 2$. (Figures 2.3, Figure 2.4). Each of the figures consists of plots of risk ratios, as functions of $\Delta = \theta_1 - \theta_2$ with the benchmark $\hat{q}_{mre}$ as the reference point. The variances are set equal to 1, except for Figure 2 which highlights the effect of varying σ_2^2.

Figure 2.1 illustrates the effectiveness of variance expansion (Corollary 2.3.15), as well as the dominance finding of Theorem 2.4.26. More precisely, the Figure relates to Example 2.3.1 where $\hat{q}_{mle}$ is improved by the variance expansion version $\hat{q}_{mle,2}$, which belongs both to the subclass of dominating densities $\hat{q}_{mle,c}$ as well as to the complete subclass of such predictive densities. The gains are impressive ranging from a minimum of about 8% at $\Delta = 0$ to a supremum value of about 44% for $\Delta \to \infty$. Moreover, the predictive density $\hat{q}_{mle,2}$ also dominates $\hat{q}_{mre}$ by duality, but the gains are more modest. Interestingly, the penalty of failing to expand is more severe than the penalty for using an inefficient *plug-in* estimator of the mean. In accordance with Theorem 2.4.26, the Bayes predictive density $\hat{q}_{\pi_{U,A}}$ improves uniformly on $\hat{q}_{mre}$ except at $\Delta = 0$ where the risks are equal. As well, $\hat{q}_{\pi_{U,A}}$ compares well to $\hat{q}_{mle,2}$, except for small Δ, with $R(\theta, \hat{q}_{mle,2}) \leq R(\theta, \hat{q}_{\pi_{U,A}})$ if and only if $\Delta \leq \Delta_0$ with $\Delta_0 \approx 0.76$.

Figure 2.2 compares the efficiency of the predictive densities $\hat{q}_{\pi_{U,A}}$ and $\hat{q}_{mre}$ for varying σ_2^2. Smaller values of σ_2^2 represent more precise estimation of θ_2 and translates to a tendency for the gains offered by $\hat{q}_{\pi_{U,A}}$ to be greater for smaller σ_2^2; but the situation is slightly reversed for larger Δ.

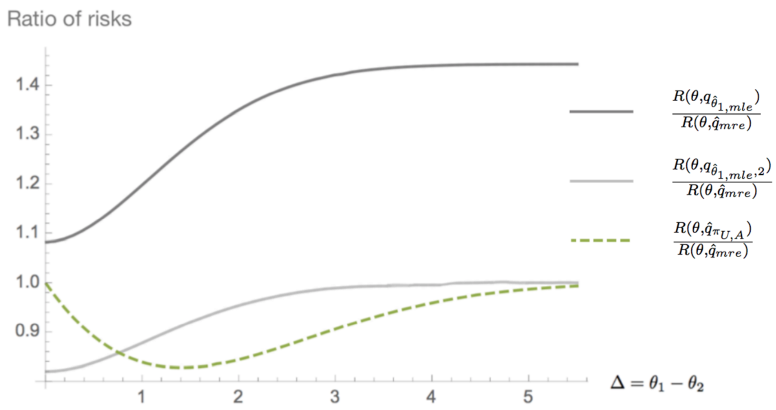

FIGURE 2.1: Kullback-Leibler risk ratios for $p = 1$, $A = [0, \infty)$, and $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2 = 1$

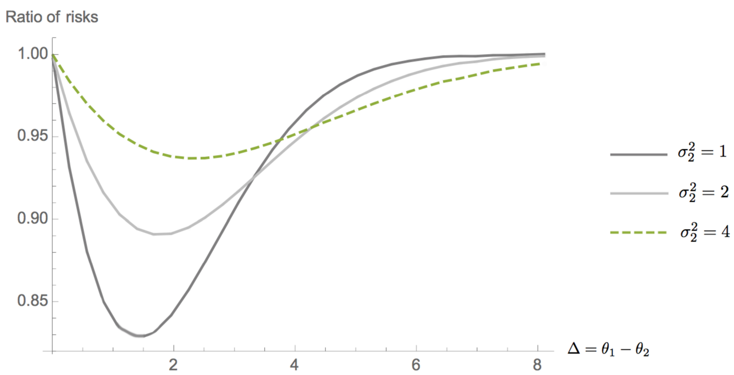

FIGURE 2.2: Kullback-Leibler risk ratios for $p = 1$, $A = [0, \infty)$, $\sigma_1^2 = \sigma_Y^2 = 1$ and $\sigma_2^2 = 1, 2, 4$

Figures 2.3 and 2.4 compare the same estimators as in Figure 2.1, but they are adapted to the restriction to compact interval. Several of the features of Figure 2.1 are reproduced with the noticeable inefficiency of $\hat{q}_{mle}$ compared to both $\hat{q}_{mle,2}$ and $\hat{q}_{\pi_{U,A}}$. For the larger parameter space (i.e. $m = 2$), even $\hat{q}_{mre}$ outperforms $\hat{q}_{mle}$ as in Figure 2.1, but the situation is reversed for $m = 1$ where the efficiency of better point maximum likelihood estimates plays a more important role. The Bayes performs well, dominating $\hat{q}_{mre}$ in accordance with Theorem 2.4.27, especially for small of moderate Δ, and even improving on $\hat{q}_{mle,2}$ for $m = 1$. Finally, we have extended the plots outside the parameter space which is useful for assessing performance for slightly incorrect specifications of the additional information.

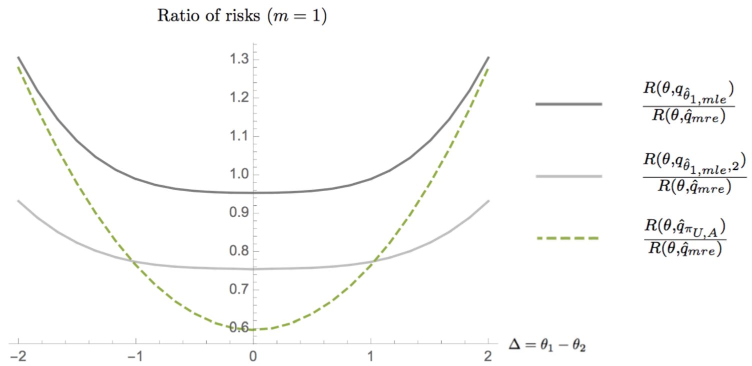

FIGURE 2.3: Kullback-Leibler risk ratios for $p = 1$, $A = [-1, 1]$, and $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2 = 1$

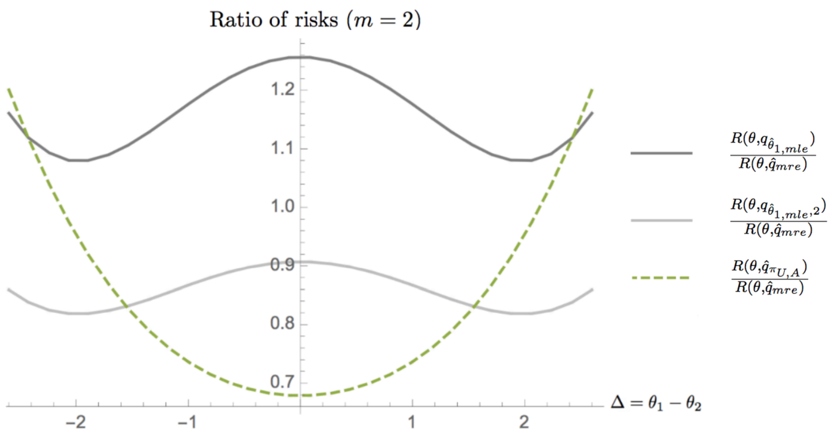

FIGURE 2.4: Kullback-Leibler risk ratios for $p = 1$, $A = [-2, 2]$, and $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2 = 1$

2.6 Concluding remarks

For multivariate normal observable $X_1 \sim N_p(\theta_1, \sigma_1^2 I_p)$, $X_2 \sim N_p(\theta_2, \sigma_2^2 I_p)$, we have provided findings concerning the efficiency of predictive density estimators $Y_1 \sim$

$N_p(\theta_1, \sigma_1^2 I_p)$ with the added parametric information $\theta_1 - \theta_2 \in A$. Several findings provide improvements on benchmark predictive densities, such those obtained as *plug-in's*, as maximum likelihood, or as minimum risk equivariant. The results range over a class of $\alpha-$divergence losses, different settings for A, and include Bayesian improvements for reverse Kullback-Leibler and Kullback-Leibler losses. The various techniques used lead to novel connections between different problems, which is also of interest as, for instance, described following both Proposition 2.3.19 and Proposition 2.3.21-Remark 2.3.22.

Although the Bayesian dominance results for Kullback-Leibler loss for $p = 1$ extend to the rectangular case with $\theta_{1,i} - \theta_{2,i} \in A_i$ for $i = 1, \ldots, p$ and the $A_i's$ either lower bounded, upper bounded, or bounded to intervals $[-m_i, m_i]$ (since the Kullback-Leibler divergence for the joint density of Y factors and becomes the sum of the marginal Kullback-Leibler divergences, and that the posterior distributions of the $\theta_{1,i}$'s are independent), a general Bayesian dominance result of $\hat{q}_{\pi_{U,A}}$ over $\hat{q}_{mre}$, is lacking and would be of interest. Finally, comparisons of predictive densities for the case of homogeneous, but unknown variance (i.e., $\sigma_1^2 = \sigma_2^2 = \sigma_Y^2$), is equally of interest.

Appendix

Predictive density estimation under reverse Kullback-Leibler loss

The objective of this part is two-fold. First, we present a quite general result which stipulates that Bayes predictive density estimators are always *plug-in* densities in an exponential family set-up with, or without, additional information. Such a result was obtained by Yanagimoto and Ohnishi (2009). We provide an extension for problems with additional information and we seek to give more prominence to Yanagimoto and Ohnishi's wonderful result. Secondly, applications of Theorems 2.6.34 and 2.6.35 yield part **(b)** of Lemma 2.2.3 and the reverse Kullback-Leibler part of Lemma 2.3.16.

Consider the exponential family model densities, with respect to $\sigma-$finite measures μ_1 and μ_2, under canonical form

$$
\begin{aligned}
X|\eta &\sim p_\eta(x) = h_1(x)\exp\{{\eta_1}^T s_1(x) + {\eta_2}^T s_2(x) - c_1(\eta)\}, \\
Y_1|\eta_1 &\sim q_{\eta_1}(y) = h_2(y_1)\exp\{{\eta_1}^T t_1(y_1) - c_2(\eta_1)\}\,,
\end{aligned} \tag{2.32}
$$

where $X = (X_1, X_2)^T$, $\eta = (\eta_1, \eta_2)^T$, and η_1, η_2, $s_1(x)$, $s_2(x)$, $t_1(y_1)$ are vectors of dimension p. In this set-up, we assume that X and Y_1 are independently distributed given η, η_1 is a common parameter, and we seek a predictive density for Y_1 based on X and with the additional information $\eta_1 - \eta_2 \in A$. We thus consider predictive densities $\hat{q}(\cdot; X)$ for Y_1 and their performance as evaluated by reverse Kullback-Leibler loss

$$
L(\eta_1, \hat{q}) = \int \hat{q}(y_1)\log\left(\frac{\hat{q}(y_1)}{q_{\eta_1}(y_1)}\right) d\mu_2(y_1)\,, \tag{2.33}
$$

and corresponding risk

$$
R(\eta, \hat{q}) = \int\int p_\eta(x)\hat{q}(y_1; x)\log\left(\frac{\hat{q}(y_1; x)}{q_{\eta_1}(y_1)}\right) d\mu_2(y_1)\, d\mu_1(x)\,.
$$

A *plug–in* estimator for the density q_{η_1} is simply of the form $q_{\hat{\eta}_1(X)}$. For Kullback-Leibler loss, obtained by switching q_{η_1} and $\hat{q}$ in (2.33), *plug-in* density estimators are not compatible with Bayesianity and can be quite inefficient in terms of Kullback-Leibler risk, as seen above in Lemma 2.3.12 for normal models. However, for reverse Kullback-Leibler loss, the situation is the opposite, and universally so for the exponential family set-up above as shown in Theorem 2.6.33. Furthermore, the *plug–in* estimator is the posterior expectation of η_1. This holds regardless of the prior on η (including cases where $\eta_1 - \eta_2 \in A$) and the particular forms of p_η and q_{η_1}. This was observed and exploited for normal models by Maruyama and Strawderman (2012).

The second observation made below concerns the frequentist risk of *plug–in* densities. Indeed, reverse Kullback-Leibler loss (among others) for a *plug-in* estimate becomes simply a measure of distance between the densities q_{η_1} and $q_{\hat{\eta}_1}$, otherwise known as intrinsic loss (e.g. Robert, 1996). For exponential families, as noted by Brown (1986, Proposition 6.3), such a distance has a simple and appealing form. Here, it leads to a representation, for both *plug–in* and thus Bayes predictive density estimators, of the reverse Kullback-Leibler risk in terms of the point estimate risk performance of the same *plug–in* estimator with respect to a dual loss.

The following representation of a Bayes predictive density estimator under reverse Kullback-Leibler is well known (e.g., Corcuera and Giummolè, 1999), but we provide a short presentation for completeness.

Lemma 2.6.33. *For estimating q_{η_1} under reverse Kullback-Leibler loss and based on X as in (2.32), the Bayes predictive density estimator is $\hat{q}_\pi(y_1;x) \propto \exp\{E(\log q_{\eta_1}(y_1)|x)\}$.*

Proof. For an estimator $\hat{q}$ and denoting G_x as the posterior c.d.f. of η, the expected posterior loss may be expressed as:

$$
\begin{aligned}
E\left(L(\eta_1,\hat{q})|x\right) &= \int\{\int \hat{q}(y_1)\left(\log\hat{q}(y_1) - \log q_{\eta_1}(y_1)\right)\,d\mu_2(y_1)\,\}\,dG_x(\eta) \\
&= \int \hat{q}(y_1)\,\{\,\log\hat{q}(y_1)\; - \;E(\log q_{\eta_1}(y_1)|x)\,\}\,d\mu_2(y_1) \\
&= \log c + \int \hat{q}(y_1)\,\{-\log(\frac{\hat{q}_\pi(y_1;x)}{\hat{q}(y_1)})\}\,d\mu_2(y_1)\,, \qquad (2.34)
\end{aligned}
$$

where $\hat{q}_\pi(y_1;x) = c\,\exp\{E(\log q_{\eta_1}(y_1)|x)\}$. Using Jensen's inequality applied to $-\log$, we obtain indeed from (2.34), for all estimators $\hat{q}$,

$$E\left(L(\eta_1,\hat{q}\,)|x\right) \geq \log\;c - \log\int \hat{q}_\pi(y_1;x)\,d\mu_2(y_1) = \log\;c = E\left(L(\eta_1,\hat{q}_\pi)|x\right). \quad \square$$

The following representation applies with or without the additional information provided by the constraint $\eta_1 - \eta_2 \in A$, with the additional information case representing an extension of Yanagimoto and Ohnishi's result.

Theorem 2.6.34. *For model (2.32), reverse Kullback-Leibler loss, a prior measure π for η such that the posterior distribution and expectation exists, the Bayes predictive density estimate $\hat{q}_\pi(\cdot;x)$ is the* plug–in *density estimate $q_{\hat{\eta}_1}(\cdot;x)$, with $\hat{\eta}_1(x) = E_\pi(\eta_1|x)$ the posterior expectation of η_1.*

Proof. Using Lemma 2.6.33, we obtain

$$
\begin{aligned}
\hat{q}_\pi(y_1;x) &\propto \exp\{E(\log q_{\eta_1}(y_1)|x)\} \\
&\propto h_2(y_1)\exp\{E(\eta_1^T t_1(y_1) - c_2(\eta_1)|x)\} \\
&\propto h_2(y_1)\exp\{(\eta_1^T t_1(y_1) - c_2(E(\eta_1)|x))\},
\end{aligned}
$$

which matches indeed the *plug-in* density $q_{\hat{\eta}_1}(\cdot;x)$ with $\hat{\eta}_1(x) = E_\pi(\eta_1|x)$. $\square$

Theorem 2.6.35. *For model (2.32), the reverse Kullback-Leibler frequentist risk of the* plug–in *density* $q_{\hat{\eta}_1}(\cdot;X)$ *is equivalent to the frequentist risk for estimating* η_1 *based on* X *under the dual point estimation loss*

$$L_{dual}(\eta_1,\hat{\eta}_1) = (\hat{\eta}_1 - \eta_1)^T \, \mathbb{E}_{\hat{\eta}_1}(t(Y)) + (c_2(\eta_1) - c_2(\hat{\eta}_1)).$$

Proof. For the plug–in density estimator, we have

$$\begin{aligned} L_{dual}(\eta_1,\hat{\eta}_1) &= \int q_{\hat{\eta}_1}(y_1) \log \frac{q_{\hat{\eta}_1}(y_1)}{q_{\eta_1}(y_1)} \, d\mu_2(y_1) \\ &= \int q_{\hat{\eta}_1}(y_1) \{(\hat{\eta}_1 - \eta_1)^T \, t(y_1) + (c_2(\eta_1) - c_2(\hat{\eta}_1))\} \, d\mu_2(y_1) \\ &= (\hat{\eta}_1 - \eta_1)^T \, \mathbb{E}_{\hat{\eta}_1} t(Y_1) + (c_2(\eta_1) - c_2(\hat{\eta}_1)) \, , \end{aligned}$$

which leads to the result. □

Example 2.6.5. *For the multivariate normal model (2.1), the last two theorems apply as examples of model (2.32) with* $\eta_1 = \theta_1$, $\eta_2 = \theta_2$, $c_2(\eta_1) = \frac{\|\eta_1\|^2}{2\sigma_Y^2}$, $t(y_1) = \frac{y_1}{\sigma_Y^2}$. *Theorem 2.6.34 yields the Bayes predictive density given in (2.5), while Theorem 2.6.35 yields the dual loss* $L_{dual}(\eta_1,\hat{\eta}_1) = (\hat{\eta}_1 - \eta_1)^T \, \mathbb{E}_{\hat{\eta}_1}(\frac{Y_1}{\sigma_Y^2}) + \frac{\|\eta_1\|^2}{2\sigma_Y^2} - \frac{\|\hat{\eta}_1\|^2}{2\sigma_Y^2} = \frac{\|\hat{\eta}_1 - \eta_1\|^2}{2\sigma^2}$*, as stated in Lemma 2.3.16.*

Acknowledgments

Author Marchand gratefully acknowledges the research support from the Natural Sciences and Engineering Research Council of Canada. We are grateful to Bill Strawderman for useful discussions, namely on the developments for reverse Kullback-Leibler loss.

Bibliography

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547-554.

Arellano-Valle, R.B., Branco, M.D., & Genton, M.G. (2006). A unified view on skewed distributions arising from selections. *Canadian Journal of Statistics*, **34**, 581-601.

Arnold, B.C. & Beaver, R.J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting (with discussion). *Test*, **11**, 7-54.

Arnold, B.C., Beaver, R.J., Groeneveld, R.A., & Meeker, W.Q. (1993). The non-truncated marginal of a truncated bivariate normal distribution. *Psychometrika*, **58**, 471-488.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171-178.

Blumenthal, S. & Cohen, A. (1968). Estimation of the larger translation parameter. *Annals of Mathematical Statistics*, **39**, 502-516.

Brandwein, A.C. & Strawderman, W.E. (1980). Minimax estimation of location parameters for spherically symmetric distributions with concave loss. *Annals of Statistics*, **8**, 279-284.

Brown, L.D., George, E.I., & Xu, X. (2008). Admissible predictive density estimation. *Annals of Statistics*, **36**, 1156-1170.

Brown, L.D. (1986). *Foundations of Exponential Families.* IMS Lecture Notes, Monograph Series **9**, Hayward, California.

Cohen, A. & Sackrowitz, H. B. (1970). Estimation of the last mean of a monotone sequence. *Annals of Mathematical Statistics*, **41**, 2021-2034.

Corcuera, J. M. & Giummolè, F. (1999). A generalized Bayes rule for prediction. Scandinavian Journal of Statistics, **26**, 265-279.

Csiszàr, I. (1967). Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. **2**, 299-318.

Dunson, D.B. & Neelon, B. (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics*, **59**, 286-295.

Fourdrinier, D. & Marchand, É. (2010). On Bayes estimators with uniform priors on spheres and their comparative performance with maximum likelihood estimators for estimating bounded multivariate normal means. *Journal of Multivariate Analysis*, **101**, 1390-1399.

Fourdrinier, D., Marchand, É., Righi, A. & Strawderman, W.E. (2011). On improved predictive density estimation with parametric constraints. *Electronic Journal of Statistics*, **5**, 172-191.

George, E. I., Liang, F. & Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *Annals of Statistics*, **34**, 78-91.

Ghosh, M., Mergel, V. & Datta, G. S. (2008). Estimation, prediction and the Stein phenomenon under divergence loss. *Journal of Multivariate Analysis*, **99**, 1941-1961.

Gupta, R.C. & Gupta, R.D. (2004). Generalized skew normal model. *Test*, **13**, 501-524.

Hartigan, J. (2004). Uniform priors on convex sets improve risk. *Statistics & Probability Letters*, **67**, 285-288.

Hwang, J. T. G. & Peddada, S. D. (1994). Confidence interval estimation subject to order restrictions. *Annals of Statistics*, **22**, 67-93.

Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, **88**, 859-864.

Kubokawa, T. (2005). Estimation of bounded location and scale parameters. *Journal of the Japanese Statistical Society*, **35**, 221-249.

Kubokawa, T., Marchand, É., & Strawderman, W.E. (2015). On predictive density estimation for location families under integrated squared error loss. *Journal of Multivariate Analysis*, **142**, 57-74.

Kubokawa, T., Marchand, É. & Strawderman, W.E. (2017). On predictive density estimation for location families under integrated absolute value loss. *Bernoulli*, **23**, 3197-3212.

Liseo, B. & Loperfido, N. (2003). A Bayesian interpretation of the multivariate skew-normal distribution. *Statistics & Probability Letters*, **61**, 395-401.

Marchand, É., & Payandeh Najafabadi, A.T. (2011). Bayesian improvements of a MRE estimator of a bounded location parameter. *Electronic Journal of Statistics*, **5**, 1495-1502.

Marchand, É., Perron, F., & Yadegari, I. (2017). On estimating a bounded normal mean with applications to predictive density estimation. *Electronic Journal of Statistics*, **11**, 2002-2025.

Marchand, É. & Perron, F. (2001). Improving on the MLE of a bounded normal mean. *Annals of Statistics*, **29**, 1078-1093.

Marchand, É., & Strawderman, W. E. (2005). On improving on the minimum risk equivariant estimator of a location parameter which is constrained to an interval or a half-interval. *Annals of the Institute of Statistical Mathematics*, **57**, 129-143.

Marchand, É. & Strawderman, W.E. (2004). Estimation in restricted parameter spaces: A review. *Festschrift for Herman Rubin*, IMS Lecture Notes-Monograph Series, **45**, 21-44.

Maruyama, Y. & Strawderman, W.E. (2012). Bayesian predictive densities for linear regression models under $\alpha-$divergence loss: Some results and open problems. *Contemporary Developments in Bayesian analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*, Institute of Mathematical Statistics Volume Series, **8**, 42-56.

Park, Y., Kalbfleisch, J.D. & Taylor, J. (2014). Confidence intervals under order restrictions. *Statistica Sinica*, **24**, 429-445.

Robert, C.P. (1996). Intrinsic loss functions. *Theory and Decision*, **40**, 192–214.

Shao, P. Y.-S. & Strawderman, W. (1996). Improving on the mle of a positive normal mean. Statistica Sinica, **6**, 275-287.

Spiring, F.A. (1993). The reflected normal loss function. *Canadian Journal of Statistics*, **31**, 321-330.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135-1151.

van Eeden, C. & Zidek, J.V. (2001). Estimating one of two normal means when their difference is bounded. *Statistics & Probability Letters*, **51**, 277-284.

van Eeden, C. & Zidek, J.V. (2003). Combining sample information in estimating ordered normal means. *Sankhyā A*, **64**, 588-610.

van Eeden, C. (2006). *Restricted parameter space problems: Admissibility and minimaxity properties.* Lecture Notes in Statistics, **188**, Springer.

Yanagimoto, T. & Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e-mixture. *Journal of Statistical Planning and Inference*, **139**, 3064-3075.

Chapter 3

Predictive Density Estimation With Unknown Variance

3.1 Introduction

This chapter mainly focuses on obtaining efficient predictive density estimates for the following model:

$$X_i \sim \mathbb{N}_p(\theta_i, \sigma^2 I_p),\ i = 1, 2, \quad Y_1 \sim \mathbb{N}_p(\theta_1, \sigma^2 I_p), \quad S^2 \sim \sigma^2\, \chi^2_k, \text{ independent}, \tag{3.1}$$

with $k \geq 2$, $\theta_1 \in \mathbb{R}^p$, $\theta_2 \in \mathbb{R}^p$, $\sigma^2 \in \mathbb{R}_+$, and $\theta_1 - \theta_2 \in A \subseteq \mathbb{R}^p$. We consider that $\theta = (\theta_1, \theta_2)$ and σ^2 are unknown, and the objective is to obtain a predictive density estimate for Y_1.

Past research has indicated that plug–in predictive density estimators can be inefficient and improve upon. (Geisser [30], Komaki [31], Lawless and Fredette [32]). For the above normal problem with $A = \mathbb{R}^p$, Aitchison [7] proved that the plug-in predictive density estimator based on $\hat{\theta}_1(X_1) = X_1$, $\hat{\sigma}^2(X_1) = S^2$ is dominated by the MRE predictive density estimator under KL loss. The MRE predictive density estimator under KL loss is obtained by choosing the prior $\pi_0(\theta, \sigma) = 1/\sigma$ (see Lemma 3.2.6). Kato [33] proved for $p \geq 3$ that the MRE predictive density estimator is dominated under KL loss by the Bayes predictive density associated with the prior $\pi(\theta, \sigma) = \frac{1}{\sigma}\|\theta\|^{2-p}$, while Boisbunon and Maruyama [34] obtained other dominating Bayes predictive density estimators even for $p = 1, 2$.

The results of this section are original as they exploit the additional information

$\theta_1 - \theta_2 \in A$ through the added observation X_2. In Section 3.2, we expand on Bayesian posterior analysis and predictive densities, as well as interesting representations in terms of skewed Student t distributions. In Section 3.3, we consider the KL risk of plug–in predictive density estimators and we expand on a point estimation dual loss, which leads to improvements. Section 3.4 expands along similar lines for RKL loss, with the difference that Bayes predictive density estimators are necessarily plug–in densities (Yanagimoto and Ohnishi [23]; Maruyama and Strawderman [35], Example 1.4.5).

3.2 Bayes posterior analysis

3.2.1 Posterior distributions and expectations

The joint density corresponding to (X, S^2) in model (3.1), supported on $\mathbb{R}^{2p} \times \mathbb{R}_+$, is given by

$$p_{\theta,\sigma^2}(x, s^2) = \frac{(s^2)^{k/2-1}}{(2\pi\sigma^2)^p} \, \frac{\exp\left\{-\frac{1}{2\sigma^2}\left(\|x_1-\theta_1\|^2 + \|x_2-\theta_2\|^2 + s^2\right)\right\}}{(2\sigma^2)^{k/2}\Gamma(k/2)} \,. \tag{3.2}$$

We consider the prior density

$$\pi_A(\theta, \sigma^2) = \frac{1}{\sigma^2} \, \mathbb{I}_A(\theta_1 - \theta_2), \tag{3.3}$$

supported on the restricted parameter space $\theta_1 - \theta_2 \in A \in \mathbb{R}^p$. In the next lemma we provide the posterior density $\theta \,|\, x, s^2$ and marginal posterior density $\theta_1 \,|\, x, s^2$, which are needed for determining $\hat{q}_{\pi_A}$.

Lemma 3.2.1. *For model (3.1) and prior density (3.3), we have that*

(a) *the posterior density $\pi(\theta \,|\, x, s^2)$ is proportional to*

$$\left(1 + \frac{\|x_2-\theta_2\|^2}{s^2 + \|x_1-\theta_1\|^2}\right)^{-(p+k/2)} \left(1 + \frac{\|x_1-\theta_1\|^2}{s^2}\right)^{-(p+k/2)} \mathbb{I}_A(\theta_1 - \theta_2)\,. \tag{3.4}$$

(b) *the marginal posterior density is given by*

$$\pi(\theta_1 \,|\, x, s^2) \propto f_{k,x_1,\frac{s}{\sqrt{k}}}(\theta_1) \, \mathbb{P}(V \in A)\,,$$

where $f_{\nu,\xi,\tau}$ is the density of $\mathbb{T}_p(\nu = k,\, \xi = x_1, \tau = \frac{s}{\sqrt{k}})$ density and

$$V \sim \mathbb{T}_p\left(\nu = k + p, \xi = \theta_1 - x_2, \tau = \sqrt{\frac{s^2 + \|x_1 - \theta_1\|^2}{p+k}}\right). \tag{3.5}$$

Proof. **(a)** We have

$$\pi(\theta, \sigma^2 \,|\, x, s^2) \propto (\sigma^2)^{-(p+k/2+1)} \exp\left\{\frac{-t}{2\sigma^2}\right\} \mathbb{I}_A(\theta_1 - \theta_2)\,, \tag{3.6}$$

where $t = \|x_1 - \theta_1\|^2 + \|x_2 - \theta_2\|^2 + s^2$. Now, letting $z = \frac{t}{2\sigma^2}$, by integrating out σ^2, we have

$$\begin{aligned}
\pi(\theta \,|\, x, s^2) &\propto \int_0^\infty (\sigma^2)^{-(p+k/2+1)} \exp\{-\frac{t}{2\sigma^2}\} \,\mathbb{I}_A(\theta_1 - \theta_2)\, d\sigma^2 \\
&\propto t^{-(p+k/2)}\, \mathbb{I}_A(\theta_1 - \theta_2) \int_0^\infty z^{p+k/2-1} \exp\{-z\}\, dz \\
&\propto t^{-(p+k/2)}\, \mathbb{I}_A(\theta_1 - \theta_2) \\
&\propto (\|x_1 - \theta_1\|^2 + \|x_2 - \theta_2\|^2 + s^2)^{-(p+k/2)} \mathbb{I}_A(\theta_1 - \theta_2) \\
&\propto \left(1 + \frac{\|x_2 - \theta_2\|^2}{s^2 + \|x_1 - \theta_1\|^2}\right)^{-(p+k/2)} \left(1 + \frac{\|x_1 - \theta_1\|^2}{s^2}\right)^{-(p+k/2)} \mathbb{I}_A(\theta_1 - \theta_2).
\end{aligned}$$

(b) We have

$$\begin{aligned}
\pi(\theta_1 \,|\, x, s^2) &= \int\limits_{\{\theta_2 :\, \theta_1 - \theta_2 \in A\}} \pi(\theta \,|\, x, s^2)\, d\theta_2 \\
&\propto \left(1 + \frac{\|x_1 - \theta_1\|^2}{s^2}\right)^{-(p+k/2)} \int\limits_{\{\theta_2 :\, \theta_1 - \theta_2 \in A\}} \left(1 + \frac{\|x_2 - \theta_2\|^2}{s^2 + \|x_1 - \theta_1\|^2}\right)^{-(p+k/2)} d\theta_2, \\
&\propto \left(1 + \frac{\|\theta_1 - x_1\|^2}{k(s^2/k)}\right)^{-(p+k/2)} \int\limits_A \left(1 + \frac{\|t - (\theta_1 - x_2)\|^2}{s^2 + \|x_1 - \theta_1\|^2}\right)^{-(p+k/2)} dt\,,
\end{aligned}$$

by the change of variable $t = \theta_1 - \theta_2$. □

For example for $p = 1$ and $A = [0, \infty)$, the probability as defined in Lemma 3.2.1 (b) is equivalent to

$$\mathbb{P}(V > 0) = F_1\left(k + 1, \frac{\theta_1 - x_2}{\sqrt{\frac{s^2 + \|x_1 - \theta_1\|^2}{p+k}}}\right), \tag{3.7}$$

where $F_1(\nu,\,\cdot)$ is cdf of a standard Student t distribution with degrees of freedom $\nu > 1$. The following lemma (e.g. see Azzalini and Capitanio [28]) will be used below.

Lemma 3.2.2. *For $\eta \sim Gamma(a,b)$, and $a>0$, $b>0$, $c>0$, we have*

$$\mathbb{E}[\Phi_p(c\sqrt{\eta};0)] = F_p\left(2a,\, c\sqrt{\frac{a}{b}}\right)\,, \tag{3.8}$$

where Φ_p is cdf of a $\mathbb{N}_p(0,I_p)$ distribution and $F_p(\nu,\cdot)$ is cdf of $\mathbb{T}_p(\nu,0,1)$ Student t distribution with degrees of freedom ν.

Lemma 3.2.3. *For model (3.1), prior (3.3), $\eta = 1/\sigma^2$ and $u = \sqrt{\eta}(\theta_1 - x_1)$, the joint posterior density $(U,\eta)\,|\,x,s^2$ is proportional to weighted normal and weighted gamma densities:*

$$\pi(u\,|\,\eta,x,s^2) \;=\; \phi_p(u)\,\frac{\mathbb{P}(W'\in A)}{\mathbb{P}(V\in A)}, \tag{3.9}$$

$$\pi(\eta\,|\,x,s^2) \;=\; \frac{\eta^{k/2-1}e^{-s^2\eta/2}}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{k/2}}\,\frac{\mathbb{P}(V\in A)}{\mathbb{P}(V'\in A)}, \tag{3.10}$$

where $W' \sim \mathbb{N}_p(\frac{u}{\sqrt{\eta}}+x_1-x_2, I_p/\eta)$, $V \sim \mathbb{N}_p(x_1-x_2, 2/\eta I_p)$, and $V' \sim \mathbb{T}_p(k, \frac{x_1-x_2}{\sqrt{\frac{2s^2}{k}}}, 1)$.

Proof. We have

$$\pi(\theta_1,\sigma^2\,|\,x,s^2) \propto e^{-\frac{s^2}{2\sigma^2}}(\sigma^2)^{-(\frac{p+k}{2}+1)}\phi(\frac{\theta_1-x_1}{\sigma})\int_{\{\theta_2:\theta_1-\theta_2\in A\}}(\sigma^2)^{-\frac{p}{2}}\,\phi(\frac{\theta_2-x_2}{\sigma})\,d\theta_2.$$

$$\propto e^{-\frac{s^2}{2\sigma^2}}(\sigma^2)^{-(\frac{p+k}{2}+1)}\,\phi(\frac{\theta_2-x_2}{\sigma})\;\mathbb{P}(W\in A), \tag{3.11}$$

where $W \sim \mathbb{N}_p(\theta_1 - x_2,\, \sigma^2\, I_p)$. From this a change of variables

$$\pi(u\,|\,\eta,x,s^2) = \frac{\eta^{k/2-1}e^{-s^2\eta/2}}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{k/2}}\,\frac{\mathbb{P}(W'\in A)}{\mathbb{E}^U(\mathbb{P}(W'\in A))}.$$

In the above equation, the expectation is given by

$$\int \phi(u)\int_A \eta^{\frac{p}{2}}\phi\left(\eta(t-(\frac{u}{\sqrt{\eta}}+x_1-x_2))\right)\,dt\,du = \mathbb{P}(V\in A),$$

and the result in (3.9) is obtained after some algebra.
To prove (3.10), we have

$$\pi(\eta\,|\,x,s^2) = \frac{\eta^{k/2-1}e^{-s^2\eta/2}}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{k/2}}\frac{\mathbb{P}(W'\in A)}{\mathbb{E}^{\eta}(\mathbb{P}(W'\in A))}.$$

The expectation in the denominator above, is given by

$$\begin{aligned}\mathbb{E}^{\eta}[\mathbb{P}(V\in A)] &= \mathbb{E}^{T}\left(\mathbb{P}\left(Z=\sqrt{\frac{T}{2}}(v-(x_1-x_2)\in\left[A-(x_1-x_2)\sqrt{\frac{T}{2}}\right]\,|\,T=t\right)\right)\\ &= \mathbb{P}\left(\frac{Z}{\sqrt{\frac{T}{2}\frac{k}{s^2}}}\in(x_1-x_2)\sqrt{\frac{2s^2}{k}}\right),\end{aligned}$$

where $Z\sim\mathbb{N}_p(0,I_p)$, $T\sim Gamma(\frac{k}{2},\frac{s^2}{2})$, and hence $\frac{Z}{\sqrt{\frac{T}{2}\frac{k}{s^2}}}$ has standard Student t distribution with degrees of freedom k. □

The following corollary studies the joint posterior density $(U,\eta)\,|\,x,s^2$ with the specific constraint $\theta_1-\theta_2\in A=\mathbb{R}^p_+$.

Corollary 3.2.4. *Under the assumptions of Lemma 3.2.3, with constraint $A=\mathbb{R}^p_+$, we have*

$$\theta_1\,|\,\eta,x,s^2\sim\mathbb{SN}_p\left(\alpha_0=(x_1-x_2)\sqrt{\frac{\eta}{2}},\alpha_1=1,\xi=1,\tau=\frac{1}{\sqrt{\eta}}\right), \tag{3.12}$$

and,

$$\eta\,|\,x,s^2\sim\pi_{U,A}(\eta\,|\,x,s^2)=\frac{\eta^{k/2-1}e^{-s^2\eta/2}}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{k/2}}\frac{\Phi_p\left((x_1-x_2)\sqrt{\frac{\eta}{2}};0\right)}{F_p\left(k,\frac{x_1-x_2}{\sqrt{2s^2/k}}\right)}, \tag{3.13}$$

where Φ_p is cdf of a $\mathbb{N}_p(0,I_p)$ and $F_p(\nu,\cdot)$ is a cdf of a p–variate Student t distribution with degrees of freedom ν.

Proof. In (3.11), $\mathbb{P}(W\in A)$ can be replaced by $\Phi_p(\frac{\theta_1-x_2}{\sigma};0)$. Thus

$$\begin{aligned}\pi_{U,A}(\theta_1,\sigma^2\,|\,x,s^2) &\propto \frac{e^{\frac{-s^2}{2\sigma^2}}}{(\sigma^2)^{(k/2+1)}}(\frac{1}{\sigma^2})^{\frac{p}{2}}\phi_p(\frac{\theta_1-x_1}{\sigma})\Phi_p(\frac{\theta_1-x_1}{\sigma};0)\\ &\propto \Phi_p(\frac{x_1-x_2}{\sqrt{2}\tau};0)\frac{e^{\frac{-s^2}{2\sigma^2}}}{(\sigma^2)^{(k/2+1)}}\frac{(\frac{1}{\tau})^p\phi_p(\frac{\theta_1-\xi}{\tau})\Phi_p(\alpha_0+\alpha_1\frac{\theta_1-\xi}{\tau};0)}{\Phi_p(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}};0)},\end{aligned}$$

and changing the variable $\eta = 1/\sigma^2$, proves (3.12).
In addition, we can write

$$\pi_{U,A}(\eta \,|\, x, s^2) \propto \frac{\eta^{k/2-1} e^{-s^2\eta/2}}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{k/2}} \, \Phi_p\left((x_1 - x_2)\sqrt{\frac{\eta}{2}}; 0\right).$$

Now, choosing $a = k/2$, $b = s^2/2$ in (3.8), we have

$$\pi_{U,A}(\eta \,|\, x, s^2) = \frac{\eta^{k/2-1} e^{-s^2\eta/2}}{\Gamma(\frac{k}{2})(\frac{2}{s})^{k/2}} \, \frac{\Phi_p\left((x_1 - x_2)\sqrt{\frac{\eta}{2}}; 0\right)}{F_p\left(k, \frac{x_1 - x_2}{\sqrt{\frac{2s^2}{k}}}\right)},$$

establishing and completing the proof. (3.13). □

Similarly, one can consider constraint of the form $\theta_1 - \theta_2 \in A = [-m, m]^p$ in Lemma 3.2.7 leading to the following corollary.

Corollary 3.2.5. *Under the assumptions of Lemma 3.2.3, with constraint $\theta_1 - \theta_2 \in A = [-m, m]^p$, the marginal posterior distribution $\pi(\theta_1 \,|\, \eta, x, s^2)$ is given by*

$$\mathbb{SN}_p\left(\alpha_0 = (x_1 - x_2 + m)\sqrt{\frac{\eta}{2}}, \alpha_1 = 1, \alpha_2 = (x_1 - x_2 - m)\sqrt{\frac{\eta}{2}}, \xi = 1, \tau = \frac{1}{\sqrt{\eta}}\right),$$

and also,

$$\pi(\eta \,|\, x, s^2) = \frac{\eta^{k/2-1} e^{-s\eta/2}}{\Gamma(\frac{k}{2})(\frac{2}{s})^{k/2}} \, \frac{\Phi_p\left((x_1 - x_2 + m)\sqrt{\frac{\eta}{2}}; 0\right) - \Phi_p\left((x_1 - x_2 - m)\sqrt{\frac{\eta}{2}}; 0\right)}{F_p(k, \frac{x_1 - x_2 + m}{\sqrt{2s^2/k}}) - F_p(k, \frac{x_1 - x_2 - m}{\sqrt{2s^2/k}})}.$$

Proof. The proof is similar to Corollary 3.2.4 using the fact that $\mathbb{P}(W \in A) = \Phi_p(\frac{\theta_1 - x_2 + m}{\sigma}; 0) - \Phi_p(\frac{\theta_1 - x_2 - m}{\sigma}; 0)$ in (3.11). □

3.2.2 Predictive densities

The MRE predictive density estimator $\hat{q}_{mre}$, which is the generalized Bayes predictive density estimator with respect to the non–informative prior $\pi_0(\theta_1, \sigma^2) = \frac{1}{\sigma^2}$ for (θ_1, σ^2) based on (X_1, S^2) is considered as a benchmark density estimator.

Lemma 3.2.6. *For model (3.1), and KL loss, the MRE predictive density estimate for density of Y_1 is given by*

$$\hat{q}_{mre}(y_1; x_1, s^2) \sim \mathbb{T}_p\left(\nu = k, \xi = x_1, \tau = \sqrt{\frac{2s^2}{k}}\right). \tag{3.14}$$

Proof. For the non–informative prior $\pi_0(\theta_1, \sigma^2)$ we have

$$\pi(\theta_1, \sigma^2 \,|\, x_1, s^2) \propto (\sigma^2)^{-\frac{p+k}{2}-1} \exp\{\frac{-t'}{2\sigma^2}\},$$

where $t' = \|x_1 - \theta_1\|^2 + s^2$. This gives $\pi(\sigma^2 \,|\, x_1, s^2) \propto (\sigma^2)^{-(k/2+1)} \exp\{-\frac{s^2}{2\sigma^2}\}$ and recognized as a scale inverse chi–square, $SInv - \chi^2(k, \sqrt{\frac{s^2}{k}})$. Therefore we have

$$\begin{aligned}
q(y_1; x_1, s^2) &= \int_0^\infty q(y_1 \,|\, x_1, \sigma^2)\pi(\sigma^2 \,|\, x_1, s^2)\, d\sigma^2 \\
&\propto \int_0^\infty (\sigma^2)^{-p/2} \exp\{-\frac{\|y_1 - x_1\|^2}{2\sigma^2}\}(\sigma^2)^{-(k/2-1)} \exp\{-\frac{-t'}{2\sigma^2}\}\, d\sigma^2 \\
&\propto \int_0^\infty (\sigma^2)^{-\frac{p+k}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left(\|y_1 - x_1\|^2 + \frac{s^2}{2}\right)\right\} d\sigma^2 \\
&\propto t''^{-\frac{p+k}{2}} \int_0^\infty z^{\frac{p+k}{2}-1} \exp\{-z\}\, dz, \quad \text{with } t'' = \|y_1 - x_1\|^2 + \frac{s^2}{2} \\
&\propto \left(1 + \frac{2k\|y_1 - x_1\|^2}{k\, s^2}\right)^{-\frac{p+k}{2}}.
\end{aligned}$$

This is the kernel of $\mathbb{T}_p(\nu = k, \xi = x_1, \tau = \sqrt{\frac{s^2}{2k}})$ and hence the proof. □

The, next lemma enables us to express predictive density estimates of density of Y_1 under additional information as a weighted Student t distribution. These densities under some of specific constraints on parameters belong to skewed–Student t distribution varying with A (see examples 3.2.1 and 3.2.2).

Theorem 3.2.7. *For model (3.1), prior (3.3) on $A \subseteq \mathbb{R}^p$, the Bayes predictive density $\hat{q}_{\pi,A}(y_1; x, s^2)$ associated with KL loss is given by*

$$\hat{q}_{\pi,A}(y_1;\, x) \;=\; \hat{q}_{mre}(y_1;\, x_1)\, \frac{\mathbb{P}(V'' \in A)}{\mathbb{P}(V' \in A)}, \tag{3.15}$$

where $V' \sim \mathbb{T}_p(k, \frac{x_1 - x_2}{\sqrt{\frac{2s^2}{k}}})$ and $V'' \sim \mathbb{T}_p(k + p, \frac{x_1 - x_2}{\sqrt{\frac{2s^2}{k}}})$.

Proof. Set $\eta = 1/\sigma^2$ and $u = \sqrt{\eta}(\theta_1 - x_1)$. The joint density of $(U,\eta)\,|\,x,s^2$ is given by multiplication of equations (3.9) and (3.10). According to (1.2),

$$\begin{aligned}
\hat{q}_{\pi,A}(y_1;x,s^2) &= \mathbb{E}\left[q(y_1\,|\,x_1+\frac{U}{\sqrt{\eta}},\frac{1}{\eta})\,|\,x,s^2\right] \\
&= \int_{\mathbb{R}_+^p}\left(\int_{\mathbb{R}^p}(\frac{\eta}{2\pi})^{\frac{p}{2}}e^{-\frac{\eta}{2}\|y_1-x_1-\frac{u}{\sqrt{\eta}}\|^2}\frac{\phi_p(u)\mathbb{P}(W'\in A)}{\mathbb{P}(V\in A)}du\right)\pi(\eta\,|\,x,s^2)\,d\eta \\
&= \int_{\mathbb{R}_+^p}\frac{\eta^{\frac{k+p}{2}-1}e^{-\eta/2(s^2+\|y_1-x_1\|^2)}}{\mathbb{P}(V\in A)\Gamma(\frac{k}{2})(\frac{2}{s^2})^{k/2}}\int_{\mathbb{R}^p}\frac{e^{-\|u\|^2/2+\sqrt{\eta}u^T(y_1-x_1)}}{(2\pi)^{\frac{p}{2}}}\phi_p(u)\mathbb{P}(V'\in A)\,du\,d\eta \\
&= \frac{\Gamma(\frac{k+p}{2})}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{\frac{k}{2}}(2\pi)^{\frac{p}{2}}\mathbb{P}}\left(\frac{s^2}{2}+\frac{\|y_1-x_1\|^2}{4}\right)^{-\frac{k+p}{2}} \\
&\times\int_{\mathbb{R}_+^p}\eta^{\frac{k+1}{2}-1}e^{-\eta\left(\frac{s^2}{2}+\frac{\|y_1-x_1\|^2}{4}\right)}\frac{1}{\sqrt{2}}\mathbb{P}(V'\in A)\,d\eta \\
&= \frac{\Gamma(\frac{k+p}{2})}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{\frac{k}{2}}(2\pi)^{\frac{p}{2}}\mathbb{P}}\frac{1}{\sqrt{2}}\mathbb{E}^{\eta\,|\,x,s^2}\left[\mathbb{P}(W'\in A)\right],
\end{aligned}$$

where $\eta\,|\,x,s^2 \sim Gamma(\frac{k+1}{2},\,\frac{s^2}{2}+\frac{\|y_1-x_1\|^2}{4})$. Finally applying identity (3.8) to above expectation completes the proof. □

The next lemma relates to Bayes predictive density estimates associated with reverse Kullback–Leibler loss.

Lemma 3.2.8. *Under the assumptions of Theorem 3.2.7, with constraint $A = \mathbb{R}_+^p$, we have*

$$\mathbb{E}[\eta\,|\,x,s^2] = \frac{k}{s^2}\frac{F_p(k+2,\,\frac{x_1-x_2}{\sqrt{2s^2/k}})}{F_p(k,\,\frac{x_1-x_2}{\sqrt{2s^2/k}})}, \tag{3.16}$$

$$\mathbb{E}[\theta_1\eta\,|\,x,s^2] = x_1\frac{k}{s^2}\frac{F_p(k+2,\,\frac{x_1-x_2}{\sqrt{2s^2/k}})}{F_p(k,\,\frac{x_1-x_2}{\sqrt{2s^2/k}})} + \frac{1}{F_p(k,\,\frac{x_1-x_2}{\sqrt{2s^2/k}})}\frac{\Gamma(\frac{k+p}{2})}{\Gamma(\frac{k}{2})\,(2\pi)^{\frac{p}{2}}s^2} \left(1+\frac{(x_1-x_2)^2}{2s^2}\right)^{-\frac{k+p}{2}}, \tag{3.17}$$

where $F_p(\nu,\cdot)$ is cdf of a standard p–variate Student t distribution with degrees of freedom ν.

Proof. Making use of posterior density in (3.13), yields (3.16).
In order to prove (3.17), one can write

$$\begin{aligned}\mathbb{E}[\eta\,\theta_1\,|\,x,s^2] &= \mathbb{E}^{\eta\,|x,s^2}\left[\eta\,\mathbb{E}(\theta_1\,|\,\eta,x,s^2)\right]\\ &= \mathbb{E}^{\eta\,|\,x,s^2}\left[\eta\left(x_1+(\frac{1}{2\eta})^{\frac{p}{2}}\frac{\phi_p\left((x_1-x_2)\sqrt{\frac{\eta}{2}}\right)}{\Phi_p\left((x_1-x_2)\sqrt{\frac{\eta}{2}}\right)}\right)\right]\\ &= \mathbb{E}^{\eta\,|\,x,s^2}\eta+\mathbb{E}^{\eta\,|\,x,s^2}\left[\frac{\phi_p\left((x_1-x_2)\sqrt{\frac{\eta}{2}}\right)}{\Phi_p\left((x_1-x_2)\sqrt{\frac{\eta}{2}}\right)}\right],\end{aligned}$$

by replacing (3.16) in the first part in above as well as some algebra for the second part, yields the result. □

We conclude this section with predictive density examples for cases $A=\mathbb{R}_+{}^p$ and $A=[-m,m]^p$.

Example 3.2.1. *For model (3.1) and prior (3.3) with $A=\mathbb{R}_+{}^p$, the Bayes predictive density $\hat{q}_{\pi,A}(y_1;x,s^2)$ associated with KL loss is given by a*

$$\mathbb{ST}_p\left(\nu=k,\alpha_0=\sqrt{\frac{2}{3}}\frac{x_1-x_2}{\sqrt{2s^2/k}},\alpha_1=1/\sqrt{3},\,\xi=x_1,\,\tau=\sqrt{\frac{2s^2}{k}}\right).$$

or equivalently

$$\hat{q}_{\pi,A}(y_1;\,x,s^2)\;=\hat{q}_{mre}(y_1;\,x_1,s^2)\,\frac{F_p\left(k+p,\,\left(\sqrt{\frac{2}{3}}(x_1-x_2)+\frac{y_1-x_1}{\sqrt{3}}\right)\sqrt{\frac{k+1}{2s^2+(y_1-x_1)^2}}\right)}{F_p(k,\,\frac{x_1-x_2}{\sqrt{2s^2/k}})},\tag{3.18}$$

where $\hat{q}_{mre}(y_1;\,x_1,s^2)\sim\mathbb{T}_p\left(\nu=k,\xi=x_1,\tau=\sqrt{\frac{2s^2}{k}}\right)$ as given in Lemma 3.2.6 and $F_p(\nu,\cdot)$ is cdf of a standard p–variate Student t distribution with degrees of freedom ν.

To establish the above, set $U=\eta(\theta_1-x_1)$, and $\eta=\frac{1}{\sigma^2}$, one can write the joint density of $(U,\eta)\,|\,x,s^2$ as multiplication of equations (3.12) and (3.13) in Corollary

3.2.4. So, we have

$$\begin{aligned}
\hat{q}_{\pi,A}(y_1;x,s^2) &= \mathbb{E}^{(U,\eta)\,|\,x,s^2} q(y_1\,|\,x_1+\frac{U}{\sqrt{\eta}},\frac{1}{\eta}) \\
&= \int_{\mathbb{R}^p_+}\left(\int_{\mathbb{R}^p}(\frac{\eta}{2\pi})^{\frac{p}{2}}e^{-\frac{\eta}{2}\|y_1-x_1-\frac{u}{\sqrt{\eta}}\|^2}\frac{\phi_p(u)\Phi_p(\alpha_0+\alpha_1 u;0)}{\Phi_p(\frac{\alpha_0}{\sqrt{1+\alpha_1^T\alpha_1}};0)}du\right)\pi(\eta\,|\,x,s^2)\,d\eta \\
&= \int_{\mathbb{R}^p_+}\frac{\eta^{\frac{k+p}{2}-1}e^{-\eta/2(s^2+\|y_1-x_1\|^2)}}{F_p(k;\frac{x_1-x_2}{\sqrt{2s^2/k}})\Gamma(\frac{k}{2})(\frac{2}{s^2})^{k/2}}\int_{\mathbb{R}^p}\frac{e^{-\|u\|^2/2+\sqrt{\eta}u^T(y_1-x_1)}}{(2\pi)^{\frac{p}{2}}}\phi_p(u)\Phi_p(\alpha_0+\alpha_1 u;0)\,du\,d\eta \\
&= \frac{\Gamma(\frac{k+p}{2})}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{\frac{k}{2}}(2\pi)^{\frac{p}{2}}F_p(k,\frac{x_1-x_2}{\sqrt{2s^2/k}})}\left(\frac{s^2}{2}+\frac{\|y_1-x_1\|^2}{4}\right)^{-\frac{k+p}{2}} \\
&\times\int_{\mathbb{R}^p_+}\eta^{\frac{k+1}{2}-1}e^{-\eta\left(\frac{s^2}{2}+\frac{\|y_1-x_1\|^2}{4}\right)}\frac{1}{\sqrt{2}}\Phi_p\left(\sqrt{\eta}\left(\frac{x_1-x_2}{\sqrt{3}}+\frac{y_1-x_1}{\sqrt{6}};0\right)\right)d\eta \\
&= \frac{\Gamma(\frac{k+p}{2})}{\Gamma(\frac{k}{2})(\frac{2}{s^2})^{\frac{k}{2}}(2\pi)^{\frac{p}{2}}F_p(k;\frac{x_1-x_2}{\sqrt{2s^2/k}})}\frac{1}{\sqrt{2}}\mathbb{E}^{\eta\,|\,x,s^2}\left[\Phi_p\left(\sqrt{\eta}\left(\frac{x_1-x_2}{\sqrt{3}}+\frac{y_1-x_1}{\sqrt{6}}\right);0\right)\right],
\end{aligned}$$

where $\eta\,|\,x,s^2 \sim Gamma(\frac{k+1}{2},\frac{s^2}{2}+\frac{\|y_1-x_1\|^2}{4})$. Applying identity (3.8) to the above expectation yields (3.18).

Example 3.2.2. *For model (3.1), $A=[-m,m]^p$ for some $m>0$ and a uniform prior (3.3), the Bayes predictive density $\hat{q}_{\pi,A}(y_1;x,s^2)$, associated with KL loss follows is given by a*

$$\mathbb{ST}_p\left(\alpha_0=\sqrt{\frac{2}{3}}\frac{x_1-x_2+m}{\sqrt{2s^2/k}},\alpha_1=\frac{1}{\sqrt{3}},\alpha_2=\sqrt{\frac{2}{3}}\frac{x_1-x_2-m}{\sqrt{2s^2/k}}\,\xi=x_1,\,\tau=\sqrt{\frac{2s^2}{k}}\right)\text{ density.} \tag{3.19}$$

In other words,

$$\hat{q}_{\pi,A}(y_1;\,x,\,s^2) = \hat{q}_{mre}(y_1;\,x_1,\,s^2)\frac{F_p\left(k+1,L_1(x,s^2)\right)-F_p\left(k+1,L_2(x,s^2)\right)}{F_p\left(1,\frac{x_1-x_2+m}{\sqrt{2s^2/k}}\right)-F_p\left(1,\frac{x_1-x_2-m}{\sqrt{2s^2/k}}\right)}, \tag{3.20}$$

where $L_1(x,s^2)=\sqrt{\frac{2}{3}}(x_1-x_2+m)+\frac{y_1-x_1}{\sqrt{3}}\sqrt{\frac{k+1}{2s^2+\|y_1-x_1\|^2}}$, $L_2(x,s^2)=\sqrt{\frac{2}{3}}(x_1-x_2-m)+\frac{y_1-x_1}{\sqrt{3}}\sqrt{\frac{k+1}{2s^2+\|y_1-x_1\|^2}}$ and $\hat{q}_{mre}(y_1;\,x_1,\,s^2)\sim\mathbb{T}_p\left(\nu=k,\xi=x_1,\tau=\sqrt{\frac{2s^2}{k}}\right)$.

It would be interesting to compare analytically the KL risk performance of the Bayes predictive density estimator $\hat{q}_{\pi,A}$. In order to do this, consider model (3.1) with $p=1$, KL loss and prior density (3.3) on restricted parameter space $\theta_1-\theta_2\in$

A, for $A = [0, \infty)$ in estimating density of Y_1. According to Theorem 3.2.1 the difference in risk of the Bayes predictive density estimator $\hat{q}_{\pi,A}$ and the MRE predictive density estimator $\hat{q}_{mre}$ given by

$$\begin{aligned}
\Delta(\theta, \sigma^2) &= R_{KL}((\theta,\sigma^2), \hat{q}_{mre}) - R_{KL}((\theta,\sigma^2), \hat{q}_{\pi,A}) \\
&= \mathbb{E}^{X,Y_1,s^2} \log\left(\frac{\hat{q}_{\pi,A}(Y_1;X)}{\hat{q}_{mre}(Y_1;X)}\right) \\
&= \mathbb{E}^{X,Y_1,S^2} \log\left(F_{k+1}\left(\left(\sqrt{\frac{2}{3}}(X_1 - X_2) + \frac{Y_1 - X_1}{\sqrt{3}}\right)\sqrt{\frac{k+1}{2S^2 + (Y_1 - X_1)^2}}\right)\right) \\
&\quad -\mathbb{E}^{X,Y_1,S^2} \log\left(F_k\left(\frac{X_1 - X_2}{\sqrt{2S^2/k}}\right)\right). \qquad (3.21)
\end{aligned}$$

Finally based on numerical results we conjecture that $\Delta(\theta, \sigma^2) > 0$ for all $\theta_1 \geq \theta_2$ and $\sigma^2 > 0$ with equality iff $\theta_1 = \theta_2$. A similar conjecture is applies for $A = [-m, m]$.

3.3 Improving on plug–in predictive density estimators under KL loss function

In this section, we provide improvements on plug–in predictive density estimators under KL loss function. Consider plug–in predictive density estimators $q_{\hat{\theta}_1,\hat{\sigma}^2}$, defined as $\mathbb{N}_p(\hat{\theta}_1, \hat{\sigma}^2 I_p)$ densities, where $\hat{\theta}_1$ and $\hat{\sigma}^2$ are estimators of θ_1 and σ^2 based on (X, S^2) in model (3.1).

Lemma 3.3.9. *For model (3.1), the KL loss incurred by the plug-in predictive density* $q_{\hat{\theta}_1,\hat{\sigma}^2} \sim \mathbb{N}_p(\hat{\theta}_1(X,S^2), \hat{\sigma}^2(X,S^2))$ *in estimating* q_{θ_1,σ^2} *is given by*

$$L_{KL}((\theta,\sigma^2), q_{\hat{\theta}_1,\hat{\sigma}^2}) = L_E(\sigma^2, \hat{\sigma}^2) + L_Q(\theta_1, \hat{\theta}_1)/\hat{\sigma}^2\,, \qquad (3.22)$$

where

$$L_E(\sigma^2, \hat{\sigma}^2) = \frac{\hat{\sigma}^2}{\sigma^2} - \log(\frac{\hat{\sigma}^2}{\sigma^2}) - 1\,, \;\; \text{and} \;\; L_Q(\theta_1, \hat{\theta}_1) = \frac{\|\hat{\theta}_1 - \theta_1\|^2}{p}\,. \qquad (3.23)$$

Consequently, the the corresponding risk function is given by

$$\begin{aligned}
R_{KL}((\theta,\sigma^2), q_{\hat{\theta}_1,\hat{\sigma}^2}) &= \mathbb{E}^{X,S^2}\left[L_{KL}((\theta,\sigma^2), q_{\theta_1,\sigma^2})\right] \\
&= \frac{p}{2}\left[R_E(\sigma^2, \hat{\sigma}^2) + \frac{R_Q(\theta_1, \hat{\theta}_1)}{\hat{\sigma}^2}\right]\,, \qquad (3.24)
\end{aligned}$$

where $R_E(\sigma^2, \hat{\sigma}^2)$ *and* $R_Q(\theta_1, \hat{\theta}_1)$ *are the risks associated with* L_E *and* L_Q *losses respectively.*

Proof. From (1.3), we have

$$\begin{aligned}
L_{KL}((\theta,\sigma^2), q_{\hat{\theta}_1,\hat{\sigma}^2}) &= \int q_{\theta_1,\sigma^2}(y_1)\,\log\frac{q_{\theta_1,\sigma^2}(y_1)}{q_{\hat{\theta}_1,\hat{\sigma}^2}(y_1)}\,dy_1 \\
&= \int q_{\theta_1,\sigma^2}(y_1)\,\log\frac{(2\pi\sigma^2)^{-\frac{p}{2}}\exp\{-\frac{1}{2\sigma^2}\|y_1-\theta_1\|^2\}}{(2\pi\hat{\sigma}^2)^{-\frac{p}{2}}\exp\{-\frac{1}{2\hat{\sigma}^2}\|y_1-\hat{\theta}_1\|^2\}}\,dy_1 \\
&= \int q_{\theta_1,\sigma^2}(y_1)\left(\frac{p}{2}\log\frac{\hat{\sigma}^2}{\sigma^2}+\frac{1}{2}(\frac{1}{\hat{\sigma}^2}-\frac{1}{\sigma^2})\|y_1-\theta_1\|^2\right)dy_1 \\
&\quad+\frac{1}{2\hat{\sigma}^2}\|\theta_1-\hat{\theta}_1\|^2 \\
&= \frac{p}{2}\left\{\frac{\sigma^2}{\hat{\sigma}^2}-\log\frac{\sigma^2}{\hat{\sigma}^2}-1+\frac{\|\theta_1-\hat{\theta}_1\|^2}{p\,\hat{\sigma}^2}\right\}.
\end{aligned}$$

Also, equation (3.24) is a direct consequence of (3.23). □

In fact, loss function L_E in (3.23) is the entropy loss while $L_Q/\hat{\sigma}^2$. It rewards accurate estimation of θ_1 and over estimation of σ^2, which hence leads to the preference over estimators with inflated variance. Furthermore, loss function $L_E(\sigma^2, \hat{\sigma}^2)$ penalizes more on overestimating since for $f(t) = t - \log t - 1$, we have $f(c) < f(\frac{1}{c})$ for $c < 1$. Hence when the variance is known $L_E(\hat{\sigma}^2, \sigma^2) = 0$, and hence $L_{KL}((\theta, \sigma^2), \hat{q}) = \frac{\|\theta_1 - \hat{\theta}_1\|^2}{2\sigma^2}$.
In addition, setting $\hat{\theta}_1(X) = X_1$, in Lemma 3.3.9, implies $R_{KL}((\theta, \sigma^2), q_{\hat{\theta}_1,\hat{\sigma}^2}) = \frac{\mathbb{E}\|X_1-\theta_1\|^2}{2\sigma^2} = \frac{p}{2}$, which is a constant.

Lemma 3.3.10. *For model (3.1), under loss function* L_E*, and among estimators of* $a\sigma^2$ *of the form* $\hat{\sigma}^2(X) = c\,S^2$, $c > 0, a > 0$*, the optimal choice is* $c_{opt} = \frac{a}{k-2}$*, for* $k > 2$*. Furthermore* $\hat{\sigma}_1^2(X) = c_1\,S^2$ *dominates* $\hat{\sigma}_0^2(X) = c_0\,S^2$*, iff* $c_{opt} > c_1 > c_0$*. The second part is the direct consequence of the definition of* c_{opt}*.*

Proof. Since $S^2 \sim \sigma^2\,\chi_k^2$, we have

$$\begin{aligned}
R_E((\theta,\sigma^2), q_{\hat{\theta}_1,\hat{\sigma}^2}) &= \mathbb{E}^X\left(\frac{a\,\sigma^2}{c\,S^2}\right) - \mathbb{E}^X\left(\log\frac{a\,\sigma^2}{c\,S^2}\right) - 1 \\
&= \frac{a}{c(k-2)} - \log\frac{a}{c} + \log 2 + \psi(\frac{k}{2}) - 1,
\end{aligned}$$

which minimizes by $a/c = k - 2$ and $\psi(\cdot)$ is the digamma function. □

Remark 3.3.11. *Choosing $a = 2$ in Lemma 3.3.10, gives $c_{opt} = \frac{2}{k-2}$, and the corresponding estimator dominates all the plug–in density estimators of the form $c\,S^2$ with $c < \frac{2}{k-2}$ such as the MRE of σ^2 ($c = \frac{1}{k-2}$) and the unbiased estimator of σ^2 ($c = \frac{1}{k}$).*

Lemma 3.3.12. *For model (3.1), under KL loss, consider estimate of (θ_1, σ),*

$$\left(\hat{\theta}_1(x) = x_1 + \frac{s^2}{k} g(x_1),\, \hat{\sigma}^2(x) = b\, s^2\right),$$

then if $\mathbb{E}^X\left(\|g(X_1)\|^2 + 2\, div\, g(X_1)\right) \le 0$, and consequently $(\hat{\theta}(x), \hat{\sigma}^2(x))$ improves on $(\hat{\theta}'(x), \hat{\sigma}'^2(x))$ under L_E in (3.23), therefore, the plug–in density estimate $\mathbb{N}(\hat{\theta}(x), \hat{\sigma}^2(x)\, I_p)$ dominates $\mathbb{N}(\hat{\theta}'(x), \hat{\sigma}'^2(x)\, I_p)$ under KL loss function.
In addition, improvement with respect to b, in (3.22) is achieved for any choices of b varies between $\left(\frac{2}{k-2} + \frac{\underline{a}}{k}, \frac{2}{k-2} + \frac{\bar{a}}{k}\right)$, where $\bar{a} = \sup \frac{\sigma^2}{p}\left\{\mathbb{E}\left(\|g(X_1)\|^2 + 2\, div\, g(X_1)\right)\right\}$ and $\underline{a}$ is the infimum.

Proof. We have

$$\mathbb{E}^{X,S^2} \frac{\|X_1 - \theta_1 + \frac{X^2}{k} g(X_1)\|^2}{pbS^2} =$$
$$\mathbb{E}^{X,S^2}\left(\frac{\|X_1 - \theta_1\|^2 + \frac{S^4}{k^2}\|g(X_1)\|^2 + 2(X_1 - \theta_1)^T g(X_1)\frac{S^2}{k}}{pbS^2}\right) =$$
$$\frac{p\sigma^2}{pb}\frac{1}{(k-2)\sigma^2} + \frac{k\sigma^2}{pbk^2}\mathbb{E}^{X,S^2}\left(\|g(X_1)\|^2 + 2\, div g(X_1)\right) =$$
$$\frac{1}{b(k-2)} + \frac{\sigma^2}{pbk^2}\mathbb{E}^{X,S^2}\left(\|g(X_1)\|^2 + 2\, div g(X_1)\right).$$

therefore the risk function is given by

$$\mathbb{E}\left(\frac{\sigma^2}{b\sigma^2} - \log \sigma^2 + \log S^2 - 1\right) + \frac{1}{b}\left(\frac{1}{k-2} + \frac{a(\theta,\, \sigma^2)}{k}\right)$$
$$\left(-\log \sigma^2 + \mathbb{E}(\log S^2) - 1\right) + \log b + \frac{1}{b}\left(\frac{1}{k-2} + \frac{a(\theta,\, \sigma^2)}{k}\right).$$

After some algebra it is deduced that for all θ and σ^2, $b_{opt} = \frac{2}{k-2} + \frac{a(\theta,\sigma^2)}{k}$ and permits b_{opt} varies from $\frac{2}{k-2} + \frac{\underline{a}}{k}$ to $\frac{2}{k-2} + \frac{\bar{a}}{k}$. □

Remark 3.3.13. *In Lemma 3.3.12, one can examine:*

(a) *If $g(X_1) = 0$, then $b_{opt} = \frac{2}{k-2}$, since $\bar{a} = \underline{a} = 0$.*

(b) *For $p \geq 3$, the estimator $X_1 + \sigma^2 g(X_1)$ is minimax and $\bar{a} = 0$, so the choice $\frac{2}{k-2}$ will improve on for all $b > \frac{2}{k-2}$.*

(c) *For all James–Stein estimators, we have $g(X_1) = \frac{(p-2)X_1}{X_1'X_1}$ for $p > 2$ and $a = \frac{-(p-2)\sigma^2}{p}\mathbb{E}(\frac{1}{X_1'X_1})$. Consequently $\|g(X_1)\|^2 + 2\,div\,g(X_1) = \frac{-(p-2)^2}{\|X_1\|^2}$. This yields to have $\underline{a} = \frac{-(p-2)}{p}$ and $\bar{a} = 0$. Therefore b_{opt} varies from $\frac{2}{k-2} - \frac{p-2}{kp}$ to $\frac{2}{k-2}$. In addition, any choices of b less that $\frac{2}{k-2} - \frac{p-2}{kp}$ leads to an admissible estimator.*

3.4 Improving on plug–in predictive density estimators under RKL loss function

One of the reason that makes plug–in predictive density estimators are widely recognizable and appealing is that plug-in predictive density estimator is a Bayes predictive density estimator as well under RKL in the exponential families are equal thanks to Yanagimoto and Ohnishi [23], as well as by Maruyama and Strawderman [35] (See theorem 1.4.10).

Theorem 3.4.14. *Consider the plug–in predictive density estimator $q_{\hat{\theta}_1,\hat{\sigma}^2}$ and model (3.1). Under RKL loss we have*

(a) *$L_{RKL}((\theta,\sigma^2), q_{\hat{\theta}_1,\hat{\sigma}^2}) = L(\sigma^2,\hat{\sigma}^2) + L'((\theta,\sigma^2),\hat{\theta}_1)$, is the corresponding loss with*

$$L(\sigma^2,\hat{\sigma}^2) = \frac{\hat{\sigma}^2}{\sigma^2} - \log(\frac{\hat{\sigma}^2}{\sigma^2}) - 1\,, \quad \text{and} \quad L'((\theta,\sigma),\hat{\theta}_1) = \frac{\|\hat{\theta}_1 - \theta_1\|^2}{p\sigma^2}\,, \tag{3.25}$$

and the risk function given by

$$\frac{p}{2}\left(R(\sigma,\hat{\sigma}) + \frac{R_Q(\theta_1,\hat{\theta}_1)}{\sigma^2}\right).$$

(b) *For a given prior $\pi(\theta,\sigma^2)$, the Bayes predictive density under RKL loss is a plug-in predictive density $\hat{q}_{\pi,A} \sim \mathbb{N}_p\left(\frac{\hat{\eta}_1}{\hat{\eta}_2}, \frac{1}{\hat{\eta}_2}I_p\right)$, with $\hat{\eta}_1(X,S^2) = \mathbb{E}(\frac{\theta_1}{\sigma^2}|X,S^2)$ and $\hat{\eta}_2(X,S^2) = \mathbb{E}(\frac{1}{\sigma^2}|X,S^2)$.*

Proof. **(a)** We have

$$
\begin{aligned}
L_{RKL}(\theta, q_{\hat{\theta}_1,\hat{\sigma}^2}(y_1)) &= \int q_{\hat{\theta}_1,\hat{\sigma}^2} \log \frac{q_{\hat{\theta}_1,\hat{\sigma}^2}(y_1)}{q_{\theta_1,\sigma^2}(y_1)}\, dy_1 \\
&= \int q_{\hat{\theta}_1,\hat{\sigma}^2} \log \frac{(2\pi\hat{\sigma}^2)^{-\frac{p}{2}} \exp\{-\frac{1}{2\hat{\sigma}^2}\|y_1-\hat{\theta}_1\|^2\}}{(2\pi\sigma^2)^{-\frac{p}{2}} \exp\{-\frac{1}{2\sigma^2}\|y_1-\theta_1\|^2\}}\, dy_1 \\
&= \int q_{\hat{\theta}_1,\hat{\sigma}^2} \left(\frac{p}{2}\log\frac{\sigma^2}{\hat{\sigma}^2} + \frac{1}{2}(\frac{1}{\sigma^2}-\frac{1}{\hat{\sigma}^2})\|y_1-\theta_1\|^2\right) dy_1 \\
&\quad + \frac{1}{2\hat{\sigma}^2}\|\theta_1-\hat{\theta}_1\|^2 \\
&= \frac{p}{2}\left\{\log\frac{\sigma^2}{\hat{\sigma}^2} + \frac{\hat{\sigma}^2-\sigma^2}{\sigma^2} + \frac{\|\theta_1-\hat{\theta}_1\|^2}{p\sigma^2}\right\} \\
&= L(\sigma^2,\hat{\sigma}^2) + L'(\theta_1,\hat{\theta}_1).
\end{aligned}
$$

(b) Since L in (3.25) does not depend on $\hat{\theta}_1$, so the Bayes estimate of the θ_1, is obtained by minimizing the posterior risk related to L', i.e, $\hat{\eta}_1(X,S^2) = \mathbb{E}(\frac{\theta_1}{\sigma^2}|X,S^2)$. Similarly the minimizer of the posterior risk associated with L is obtained by $\hat{\eta}_2(X,S^2) = \mathbb{E}(\frac{1}{\sigma^2}|X,S^2)$. This completes the proof.

□

Corollary 3.4.15. *For any estimator $\hat{\sigma}$ of σ^2, as well as estimators δ and δ' of θ_1, the plug–in predictive density estimator $q_{\delta_1,\hat{\sigma}^2}$ dominates $q_{\delta_1,\hat{\sigma}^2}$ under RKL loss iff δ_1 dominates δ_2 under L' in (3.25).*

One can use the idea of Jafari Jozani et al. [20]'s paper to improve θ_1 (for unknown σ^2) by Setting $\delta_1 = \delta_{\phi_1}$ and $\delta_2 = \delta_{\phi_2}$ in Lemma 1.3.9 as the following.

Theorem 3.4.16. *Consider the problem of estimating θ_1 with the estimator of the form $\delta_\phi(X,S^2) = Z_2 + \phi(Z_1,S^2)$, with $Z_1 = \frac{X_1-X_2}{2}$ and $Z_2 = \frac{X_1+X_2}{2}$. Set $\mu_1 = \frac{\theta_1-\theta_2}{2}$, $\mu_2 = \frac{\theta_1+\theta_2}{2}$ and $W = \frac{S_1^2+S_2^2}{2}$. Then we have*

(a) *The frequentist risk of δ_ϕ is given by*

$$R((\theta_1,\theta_2,\sigma),\delta_\phi) = \frac{p}{2} + \frac{1}{\sigma^2}\mathbb{E}^{Z_1,W}\|\phi(Z_1,W)-\mu_1\|^2.$$

(b) *δ_{ϕ_1} dominates δ_{ϕ_2} under under L' in (3.25) iff ϕ_1 dominates ϕ_2 under loss $\|\phi-\mu_1\|^2/\sigma^2$. And consequently $q_{\delta_{\phi_1},\hat{\sigma}^2}$ dominates $q_{\delta_{\phi_2},\hat{\sigma}^2}$ under RKL loss iff for any estimator $\hat{\sigma}^2$ of σ^2.*

Lemma 3.4.17. *For model (3.1), $p = 1$, and uniform prior (3.3) on $A = \mathbb{R}_+$, the Bayes predictive density estimate associated with RKL is a $\mathbb{N}(x_1 + \kappa(x), \gamma(x))$ density, where*

$$\kappa(x) = \frac{1}{F_1(k+2, \frac{x_1-x_2}{\sqrt{2s^2/k}})} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\sqrt{2\pi k}} \left(1 + \frac{(x_1-x_2)^2}{2s^2}\right)^{-\frac{k+1}{2}}, \tag{3.26}$$

$$\gamma(x) = \frac{s^2}{k} \frac{F_1(k, \frac{x_1-x_2}{\sqrt{2s^2/k}})}{F_1(k+2, \frac{x_1-x_2}{\sqrt{2s^2/k}})}. \tag{3.27}$$

Proof. The proof is straightforward by virtue of Theorem 3.4.14 (b) and Lemma 3.2.8. □

Corollary 3.4.18. *As a direct consequence of Lemma 3.4.17, set $x_2 \to -\infty$, in equations (3.26) and (3.27) yields $\mathbb{N}(x_1, \frac{S^2}{k})$ as MRE predictive density estimate for density Y_1 under RKL loss function.*

3.5 Concluding remarks

This Chapter extends the line of work in Chapter 2, which seeks to improve the predictive density estimates of the normal model with the unknown variance and subject ti some restrictions on the parameter space under the KL and RKL loss. We have shown that the Bayes predictive density estimator of for the distribution of future observation under these situations belongs to the class of weighted Student t distribution. More specifically, we studied the restrictions $\theta_1 - \theta_2 \in A$ and $\theta_1 - \theta_2 \in [-m, m]^p$ lead to the predictive density estimators form skew-Student t densities corresponding to Definitions 1.5.19 and 1.5.20 respectively.
However, dominance results of Bayes predictive density estimator over MRE under KL loss had been conjectured in this chapter but they can be verified in the following figures.
Figure 3.1, and 3.2 present the risk ratio of the Bayes and MRE predictive density estimators for $k = 3$ in $\Delta = (\theta_1 - \theta_2)/\sigma$ based on restricted parameter $A = [0, \infty)$ and $A = [-6, 6]$ respectively. For both graphs, we have about 12% improvement in risk function.

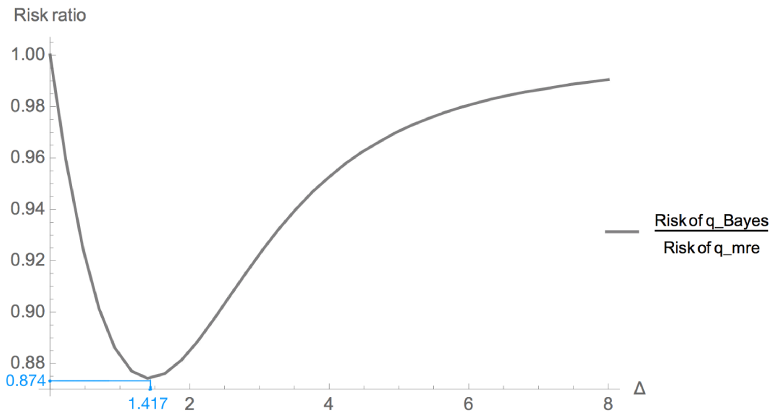

FIGURE 3.1: Risk ratio of the Bayes and MRE predictive density estimator with $k = 3$ and $A = [0, +\infty)$.

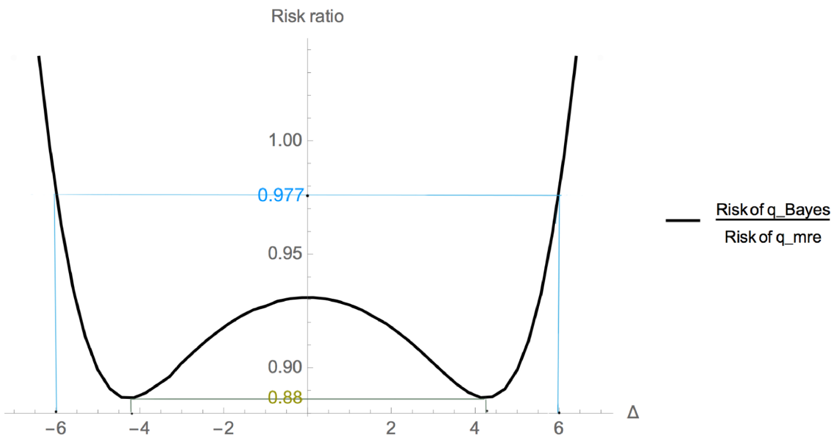

FIGURE 3.2: Risk ratio of the Bayes and the MRE predictive density estimator with $k = 3$ and $A = [-6, 6]$.

In addition, we studied comprehensively the plug–in predictive density estimators under KL and RKL and try to improve on the MRE estimator using techniques appear in Kortbi and Marchand [21] and Jafari Jozani et al.[20] borrowing from point estimation problem.

Chapter 4

Density Ratio Estimation

4.1 Introduction

In this chapter, we establish a general result for a Bayesian estimation of a ratio of two exponential family densities. The estimation of the ratio of two densities appears in several contexts, and has attracted much attention in recent years. Sugiyama et al. ([36], [37]) considered such a problem in statistical data analysis. In non-stationarity adaptation, Shimodaira [38], Sugiyama and Müller [39], Sugiyama et al. [40], Quiñonero-Candela et al. [41], and Bickel et al. [42] discussed the density ratio estimation (DRE) in multitask learning, while Hido et al. ([43], [44]) have proposed DRE as a method for statistical outlier detection.
DRE has been utilized in many applications such as: information theory, variable selection, dimension reduction, causal inference, conditional density estimation, clustering, probabilistic classification, two-sample testing, change point detection, independent component analysis, among others. Noticeably, most of the references in this area are very recent and limited to a last decade. The majority of previous studies has adapted to non-parametric and semi-parametric approaches. Here we address cases where the density belongs to the (parametric) exponential family.

4.2 Bayesian density ratio estimation

In this section, we obtain a representation for Bayesian estimators of a ratio of two exponential family densities.

One of the simplest approaches is estimating each density (say, a Bayes or plug-in predictive density estimators) separately and then forming the ratio of them. Various approaches to estimate such a ratio include: the moment mutually approach (Gretton et al. [45]), the probabilistic classification approach (Qin [46], Cheng [47]), density matching approach (Sugiyama [48]), and density-ratio fitting (Kanamori et al. [49]).
Suppose that X and Y are independently distributed with densities from exponential families defined with respect to σ–finite as follows.

$$\begin{aligned} X|\,\eta_1 \sim p_{\eta_1}(x) = h_1(x) \exp\left\{\eta_1^T\, s_1(x) - c_1(\eta_1)\right\}, \\ Y|\,\eta_2 \sim q_{\eta_2}(y) = h_2(y) \exp\left\{\eta_2^T\, s_2(y) - c_2(\eta_2)\right\}, \end{aligned} \tag{4.1}$$

where η_i are natural parameters, $s_i(\cdot)$ are sufficient statistics for $i = 1, 2$. Consider the problem of estimating the density ratio

$$r_\eta(t) = \frac{p_{\eta_1}(t)}{q_{\eta_2}(t)},$$

at some fixed point t, where $\eta = (\eta_1, \eta_2)$. A plug-in density ratio estimate is of the form $\hat{r}(t; x, y) = \frac{p_{\hat{\eta}_1(x)}(t)}{\hat{q}_{\hat{\eta}_2(y)}(t)}$, where $\hat{\eta}_1(x)$ and $\hat{\eta}_2(y)$ are estimates of parameters of $p_{\eta_1}(t)$ and $p_{\eta_2}(t)$ respectively. We show that such a estimator can be derived in a Bayesian framework in some cases. Here is a general Bayesian representation for squared log error loss.

Theorem 4.2.1. *For model (4.1), the Bayes density ratio estimate of r_η associated with loss $L(\hat{r}, r_\eta) = (\log \frac{\hat{r}}{r_\eta})^2$, and prior distribution $\pi(\eta)$, is given by*

$$\hat{r}_\pi(t; x, y) = \hat{r}(t; x, y)\, H(x, y),$$

where $\hat{r}(t; x, y) = \frac{p_{\hat{\eta}_1(x)}(t)}{\hat{q}_{\hat{\eta}_2(y)}(t)}$ is a plug–in DRE and

$$H(x, y) = \frac{\exp\left\{c_1(\,\mathbb{E}(\eta_1\,|\,x)) - \mathbb{E}(c_1(\,\eta_1\,|\,x))\right\}}{\exp\left\{c_2(\,\mathbb{E}(\eta_2\,|\,y)) - \mathbb{E}(c_2(\,\eta_2\,|\,y))\right\}}, \tag{4.2}$$

is a correction factor.

Proof. We have, $\log \hat{r}_\pi(t;x,y) = \mathbb{E}(\log r_\eta(t)|\, x,y)$, or equivalently, $\hat{r}_\pi(t;x,y) = \exp\{\mathbb{E}(\log r_\eta(t)|\, x,y)\}$. Therefore

$$\begin{aligned}
\hat{r}_\pi(t;x,y) &= \exp\{\mathbb{E}(\log p_{\eta_1}(t)|x) - \mathbb{E}(\log q_{\eta_2}(t)|y)\} \\
&= \frac{\exp\{\mathbb{E}(\log p_{\eta_1}(t)|x)\}}{\exp\{\mathbb{E}(\log q_{\eta_2}(t)|\, y)\}} \\
&= \frac{h_1(t)\, \exp\left\{\mathbb{E}(\eta_1|x)^T s_1(t) - c_1(\mathbb{E}(\eta_1|x))\right\}}{h_2(t)\, \exp\{\mathbb{E}(\eta_2|y)^T s_2(t) - c_2(\mathbb{E}(\eta_2|y))\}} \\
&\times \frac{\exp\{c_1(\mathbb{E}(\eta_1|x)) - \mathbb{E}(c_1(\eta_1)|x)\}}{\exp\{c_2(\mathbb{E}(\eta_2|y)) - \mathbb{E}(c_2(\eta_2)|y)\}} \\
&= \hat{r}(t;x,y)\, H(x,y).
\end{aligned}$$

□

4.3 Examples

Some examples are considered here. Note that, in some cases, the correction factor $H(x,y)$ is constant, i.e. it does not depend on x and y.

Example 4.3.1. *In (4.1), consider $X \sim \mathbb{N}(\mu_1,\, \sigma_1^2)$ and $Y \sim \mathbb{N}(\mu_2,\, \sigma_2^2)$ with $\eta_i = \frac{\mu_i}{\sigma_i^2}$ and $c_i(\eta_i) = \eta_i^2\sigma_i^2/2$ for $i = 1,2$. Hence the correction factor is given by*

$$\begin{aligned}
H(x,y) &= \frac{\exp\left\{\frac{\sigma_1^2}{2}\,\mathbb{E}(\eta_1|x)^2 - \frac{\sigma_1^2}{2}\,\mathbb{E}^2(\eta_1|x)\right\}}{\exp\left\{\frac{\sigma_2^2}{2}\,\mathbb{E}(\eta_2|y)^2 - \frac{\sigma_2^2}{2}\,\mathbb{E}^2(\eta_2|y)\right\}} \\
&= \frac{\exp\left\{-\frac{\sigma_1^2}{2} Var(\eta_1|x)\right\}}{\exp\left\{-\frac{\sigma_2^2}{2} Var(\eta_2|y)\right\}}.
\end{aligned} \tag{4.3}$$

Assuming a prior $\mu_i \sim \mathbb{N}(\xi_i, \tau_i^2)$, for $i = 1,2$, gives

$$\begin{aligned}
\eta_1 \,|\, X = x \;&\sim \mathbb{N}\left(\frac{{\tau_1}^2 x}{{\sigma_1}^2({\sigma_1}^2 + {\tau_1}^2)} + \frac{\xi_1}{{\sigma_1}^2 + {\tau_1}^2}\,,\; \frac{{\tau_1}^2}{{\sigma_1}^2({\sigma_1}^2 + {\tau_1}^2)}\right) \\
\eta_2 \,|\, Y = y \;&\sim \mathbb{N}\left(\frac{{\tau_2}^2 y}{{\sigma_2}^2({\sigma_2}^2 + {\tau_2}^2)} + \frac{\xi_2}{{\sigma_2}^2 + {\tau_2}^2}\,,\; \frac{{\tau_2}^2}{{\sigma_2}^2({\sigma_2}^2 + {\tau_2}^2)}\right).
\end{aligned}$$

Therefore

$$\hat{r}_\pi(t;x,y) = H(x,y)\, \hat{r}(t;x,y)\,,$$

where $H(x,y) = \exp\{-\frac{\sigma_1^2\tau_2^2-\sigma_2^2\tau_1^2}{(\tau_1^2+\sigma_1^2)(\tau_2^2+\sigma_2^2)}\}$ *is constant with respect to* x *and* y*. If* $\frac{\sigma_1}{\sigma_2} = \frac{\tau_1}{\tau_2}$*, then* $H(x,y) = 1$*, which tells us that the Bayes density ratio estimator is also a plug–in density estimator where the plug–in's are posterior expectations.*

Example 4.3.2. *Suppose that* X *and* Y *are Gamma distributed with pdfs* $p_{\lambda_1}(x) = \frac{\lambda_1{}^{\alpha_1}}{\Gamma(\alpha_1)}x^{\alpha_1-1}e^{-\lambda_1 x}$ *and* $q_{\lambda_2}(y) = \frac{\lambda_2{}^{\alpha_2}}{\Gamma(\alpha_2)}y^{\alpha_2-1}e^{-\lambda_2 y}$ *respectively where* α_1 *and* α_2 *are known. We are thus in the presence of model (4.1), with* $\eta_i = \lambda_i$ *and* $c(\eta_i) = -\alpha_i \ln(\eta_i)$ *for* $i = 1, 2$*. Hence we have*

$$H(x,y) = \frac{\exp\{-\alpha_1\,(\ln(\mathbb{E}(\eta_1|x) - \mathbb{E}(\ln\eta_1|x))\}}{\exp\{-\alpha_2\,(\ln(\mathbb{E}(\eta_2|y) - \mathbb{E}(\ln\eta_2|y))\}}. \tag{4.4}$$

Setting a prior $\lambda_i \sim Gamma(\tau_i,\, \beta_i)$ *for* $i = 1, 2$ *yields posterior distributions*

$$\lambda_1\,|\,x \;\sim\; Gamma(\tau_1+\alpha_1,\, \beta_1+x),\;\; and\;\; \lambda_2\,|\,y \;\sim\; Gamma(\tau_2+\alpha_2,\, \beta_2+y).$$

By using the fact that $\mathbb{E}\,(\ln Z) = \psi(\alpha) - \ln\lambda$*, for* $Z \sim Gamma(\alpha,\lambda)$*, where* $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ *is the digamma function, we can rewrite (4.4) as*

$$H(x,y) = \frac{(\tau_2+r_2)^{\alpha_2}}{(\tau_1+\alpha_1)^{r_1}}\,\exp\{\alpha_1\,\psi(\tau_1+\alpha_1) - \alpha_2\,\psi(\alpha_2+\alpha_2)\}\,. \tag{4.5}$$

Therefore the Bayes density estimator is of the form

$$\hat{r}_\pi(t;x,y) = H(x,y)\,\hat{r}(t;x,y)\,,$$

where $H(x,y)$ *is given in (4.5) and constant with respect to* x *and* y*. Note that the Bayes estimator is the ratio of two plug–in density estimators whenever* $\alpha_1 = \alpha_2$ *and* $\tau_1 = \tau_2$*.*

Example 4.3.3. *Let* X *and* Y *are Poisson distributed with means* λ_1 *and* λ_2 *respectively. We are thus in the presence of model (4.1) with* $\ln\lambda_i = \eta_i$ *and* $c_i(\eta_i) = e^{\eta_i}$ *for* $i = 1, 2$*. Therefore*

$$H(x,y) = \frac{e^{\mathbb{E}(\eta_1|x)} - \mathbb{E}(e^{\eta_1}|x)}{e^{\mathbb{E}(\eta_2|y)} - \mathbb{E}(e^{\eta_2}|y)}.$$

Assuming priors $\lambda_i \sim Gamma(\alpha_i,\beta_i)$ *with means* α_i/β_i*, one obtains from Theorem 4.2.1*

$$H(x,y) = \exp\left\{\frac{e^{\psi(x+\alpha_1)}}{\beta_1+1} - \frac{e^{\psi(y+\alpha_2)}}{\beta_2+1}\right\}\exp\left\{\frac{y+\alpha_2}{\beta_2+1} - \frac{x+\alpha_1}{\beta_1+1}\right\}, \tag{4.6}$$

which depends on x and y in opposition to the previous examples.

Despite the usefulness of such examples, we recall that Theorem 4.2.1 is quite general for exponential family density ratios and open the door to many further applications.
We conclude with another motivation for studying the problem of estimating the ratio of densities of the form $r_\eta(t) = \frac{p_{\eta_1}(t)}{q_{\eta_2}(t)}$ for unknown η and fixed t.

Example 4.3.4. (Estimating an $\alpha-$ divergence loss between two probability densities) *Consider an α–divergence loss between two probability densities (equation 1.5), of the form*

$$L_\alpha(p_{\eta_1}, q_{\eta_2}) = \int_{\mathbb{R}^p} h_\alpha\left(\frac{p_{\eta_1}(t)}{q_{\eta_2}(t)}\right) q_{\eta_2}(t)\, dt\,, \tag{4.7}$$

with $h_\alpha(\cdot)$ as in (1.6). So, if $r(t)$ is estimated by $\hat{r}(t)$ then, α–divergence can be estimated by the expected value of $h_\alpha(\hat{r}(t))$ under $t \sim q_{\eta_2}(t)$.

4.4 Conclusion

We have established a general representation for Bayesian estimators for exponential family density ratio and provided various examples. In some cases the Bayes estimators coincide with a ratio of plug–in density estimators.

Conclusion and future work

Conclusion

In this thesis, we tried to address the fundamental question of how we can gain from additional parametric information in order to obtain effective, and sometimes better performing, predictive densities than others in the literature. We focused on a multivariate normal model and results are applicable to Kullback–Leibler and reverse Kullback–Leibler, and the class of $\alpha-$divergence loss functions.

In Chapter 2 for the multivariate normal observable $X_1 \sim \mathbb{N}_p(\theta_1, \sigma_1^2 I_p)$, $X_2 \sim \mathbb{N}_p(\theta_2, \sigma_2^2 I_p)$, we have provided findings concerning the efficiency of predictive density estimators $Y_1 \sim \mathbb{N}_p(\theta_1, \sigma_1^2 I_p)$ with the additional information $\theta_1 - \theta_2 \in A$. We provided several results of improvements on benchmark predictive densities, such those obtained as plug-in's, maximum likelihood, and minimum risk equivariant. The findings covered $\alpha-$divergence losses, different settings for A. We showed that the obtained Bayesian predictive densities also relate to skew-normal distributions, as well as new forms of such distributions. In Chapter 3, we provided Bayes predictive density estimates for the density of $Y_1 \sim \mathbb{N}_p(\theta_1, \sigma^2 I_p)$ based on $X_i \sim \mathbb{N}_p(\theta_i, \sigma^2 I_p)$, $i = 1, 2$, independent of $S^2 \sim \sigma^2 \chi_k^2$, associated with Kullback–Leibler and reverse Kullback–Leibler losses under the restriction $\theta_1 - \theta_2 \in A$. Interesting posterior and predictive density representations arise and we provided improvements on plug–in densities. We showed that the Bayes predictive density estimator under some situations belongs to a class of skew–Student t distribution. We have established dual relationships with problems for estimating (θ_1, σ^2) that can be used for generating improved predictive densities under KL and RKL losses. In Chapter 4, we established a general form for Bayesian estimators for the ratio of exponential family densities. We showed that in some cases the Bayes estimators coincide with a ratio of plug–in density estimators or assuming that X and Y are conditionally dependent.

Future work

For future work in order to improve on predictive density estimators, it would be quite interesting to obtain more elaborate Bayesian dominance results with respect to A, the loss and even the underlying model for both known and unknown variance cases. Another possibility is to consider mixture (normal) models and try to extend the results. On the other hand, one can used some ideas in survival analysis and find the predictive density estimators in the sense of missing data and additional information.

The future work in density estimation problem may include the application of the proposed method not just at a single point t, but over a acceptable range of support space.

Bibliography

[1] Leonard Vanbrabant, Rens Van De Schoot, and Yves Rosseel. Constrained statistical inference: sample-size tables for ANOVA and regression. *Frontiers in psychology*, 5, 2014.

[2] Les John Kitchen. *Exploring statistics: A modern introduction to data analysis and inference*. West Publishing Co., 1986.

[3] Brunero Liseo and Nicola Loperfido. A bayesian interpretation of the multivariate skew-normal distribution. *Statistics & Probability letters*, 61(4):395–401, 2003.

[4] Tapan K Nayak. On best unbiased prediction and its relationships to unbiased estimation. *Journal of Statistical Planning and Inference*, 84(1):171–189, 2000.

[5] Imer Csiszàr. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[6] Malay Ghosh, Victor Mergel, and Gauri Sankar Datta. Estimation, prediction and the Stein phenomenon under divergence loss. *Journal of Multivariate Analysis*, 99(9):1941–1961, 2008.

[7] James Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975.

[8] Dominique Fourdrinier, Éric Marchand, Ali Righi, and William E. Strawderman. On improved predictive density estimation with parametric constraints. *Electron. J. Statist.*, 5:172–191, 2011.

[9] Edward I George, Feng Liang, and Xinyi Xu. Improved minimax predictive densities under, Kullback-Leibler loss. *The Annals of Statistics*, 78:78–91, 2006.

[10] Lawrence D. Brown, Edward I. George, and Xinyi Xu. Admissible predictive density estimation. *Ann. Statist.*, 36(3):1156–1170, 06 2008.

[11] Fumiyasu Komaki. A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, pages 859–864, 2001.

[12] Tatsuya Kubokawa, Éric Marchand, and William E. Strawderman. On predictive density estimation for location families under integrated squared error loss. *Journal of Multivariate Analysis*, 142:57–74, 2015.

[13] Tatsuya Kubokawa, Éric Marchand, and William E. Strawderman. On predictive density estimation for location families under integrated absolute error loss. *Bernoulli*, 23(4B):3197–3212, 2017.

[14] Éric Marchand and William E. Strawderman. A unified minimax result for restricted parameter spaces. *Bernoulli*, 18(2):635–643, 2012.

[15] Constance Van Eeden. *Restricted Parameter Space Estimation Problems: Admissibility and Minimaxity Properties*, volume 188. Springer, 2006.

[16] Eric Marchand and William E. Strawderman. Estimation in restricted parameter spaces: A review. *Lecture notes-monograph series*, 4:21–44, 2004.

[17] Saul Blumenthal and Arthur Cohen. Estimation of the larger translation parameter. *The Annals of Mathematical Statistics*, 39:502–516, 1968.

[18] Arthur Cohen and Harold B Sackrowitz. Estimation of the last mean of a monotone sequence. *The Annals of Mathematical Statistics*, 41:2021–2034, 1970.

[19] Constance Van Eeden and James V Zidek. Combining sample information in estimating ordered normal means. *Sankhyā: The Indian Journal of Statistics, Series A*, 64:588–610, 2002.

[20] Éric Marchand, Mohammad Jafari Jozani, and Yogesh Mani Tripathi. *Inadmissible estimators of normal quantiles and two-sample problems with additional information*, volume 8, pages 104–116. Institute of Mathematical Statistics, 2012.

[21] Othmane Kortbi and Éric Marchand. Estimating a multivariate normal mean with a bounded signal to noise ratio under scaled squared error loss. *Sankhya A*, 2:277–299, 2013.

[22] L. D. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Hayworth, CA, USA, 1986.

[23] Takemi Yanagimoto and Toshio Ohnishi. Bayesian prediction of a density function in terms of e-mixture. *Journal of Statistical Planning and Inference*, 139(9):3064–3075, 2009.

[24] Barry C Arnold, Robert J Beaver, A Azzalini, N Balakrishnan, A Bhaumik, DK Dey, CM Cuadras, and José María Sarabia. Skewed multivariate models related to hidden truncation and/or selective reporting. *Test*, 11(1):7–54, 2002.

[25] Ramesh C Gupta and Rameshwar D Gupta. Generalized skew normal model. *Test*, 13(2):501–524, 2004.

[26] Barry C. Arnold, Robert J. Beaver, Richard A. Groeneveld, and William Q. Meeker. The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, 58(3):471–488, 1993.

[27] Adelchi Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.

[28] Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):367–389, 2003.

[29] Reinaldo B Arellano-Valle and Marc G Genton. Multivariate unified skew-elliptical distributions. *Chilean Journal of Statistics*, 1(1):17–33, 2010.

[30] Seymour Geisser. *Predictive inference*, volume 55. CRC press, 1993.

[31] Fumiyasu Komaki. On asymptotic properties of predictive distributions. *Biometrika*, 83:299–313, 1996.

[32] J. F. Lawless and Marc Fredette. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542, 2005.

[33] Kengo Kato. Improved prediction for a multivariate normal distribution with unknown mean and variance. *Annals of the Institute of Statistical Mathematics*, 61(3):531–542, 2009.

[34] Aurélie Boisbunon and Yuzo Maruyama. Inadmissibility of the best equivariant predictive density in the unknown variance case. *Biometrika*, 101(3): 733–740, 2014.

[35] Yuzo Maruyama and William E. Strawderman. Bayesian predictive densities for linear regression models under α-divergence loss: Some results and open problems. In *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*, pages 42–56. Institute of Mathematical Statistics, 2012.

[36] Masashi Sugiyama, Takafumi Kanamori, Taiji Suzuki, Shohei Hido, Jun Sese, Ichiro Takeuchi, and Liwei Wang. A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications*, 1: 183–208, 2009.

[37] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[38] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[39] Masashi Sugiyama and Klaus-Robert Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4/2005): 249–279, 2005.

[40] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[41] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[42] Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for hiv therapy screening. In *Proceedings of the 25th international conference on Machine Learning*, pages 56–63, 2008.

[43] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Inlier-based outlier detection via direct density ratio estimation. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 223–232, 2008.

[44] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336, 2011.

[45] Arthur Gretton, Alexander J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *MIT Press*, (8):131–160, 2009.

[46] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[47] Kuang Fu Cheng and Chih-Kang Chu. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604, 2004.

[48] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

[49] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445, 2009.